Improved Cross-language Information Retrieval via Disambiguation and Vocabulary Discovery

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Ying Zhang

School of Computer Science and Information Technology
RMIT University
Melbourne, Victoria, Australia

3rd August, 2006

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Ying Zhang
School of Computer Science and Information Technology
RMIT University
3rd August, 2006

Acknowledgments

This thesis is the end of my three-year journey in obtaining the PhD degree in computer science.

With a deep sense of gratitude, I wish to express my sincere thanks to my first supervisor, Dr. Phil Vines, for his continuous guidance and encouragement throughout my masters and doctoral programs. With him, I have learned about the process of research, how to write papers, and how to make presentations. Besides being an outstanding advisor he is a good friend. My heartfelt thanks are extended to my second supervisor, Professor Justin Zobel, for his great influence on my study. He taught me how to be a critical thinker and ladylike researcher.

Finally, I would like to express very special thanks to my family and friends for their love and constant support.

Credits

Portions of work published by the author during the course of this work have formed the basis for some parts of the thesis:

- "Improved Cross-language Information Retrieval via Disambiguation and Vocabulary Discovery", in *Proceedings of the 8th Australasian Document Computing Symposium* [Zhang and Vines, 2003];
- "RMIT Chinese–English CLIR at NTCIR−4", in *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization* [Zhang and Vines, 2004c];
- "Using the Web for Automated Translation Extraction in Cross-language Information Retrieval", in *Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval* [Zhang and Vines, 2004a];
- "Detection and Translation of OOV Terms Prior to Query Time", in *Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval* [Zhang and Vines, 2004b];
- "Using the Web for Translation Disambiguation (RMIT University at NTCIR−5 Chinese–English CLIR)", in *Proceedings of the 5th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization* [Zhang and Vines, 2005];
- "Chinese OOV Translation and Post-translation Query Expansion in Chinese–English Cross-language Information Retrieval", in *ACM Transaction on Asian Language Information Processing* [Zhang et al., 2005];
- "Automatic Acquisition of Chinese–English Parallel Corpus from the Web", in *Proceedings of the 28th European Conference on Information Retrieval* [Zhang et al., 2006b];
- "An Empirical Comparison of Translation Disambiguation Techniques for Chinese–English Cross-language Information Retrieval", in *Proceedings of the 3rd Asia Information Retrieval Symposium* [Zhang et al., 2006a].

Contents

List of Figures

List of Tables

Abstract

Cross-lingual information retrieval (CLIR) allows people to find documents irrespective of the language used in the query or document. This thesis is concerned with the development of techniques to improve the effectiveness of Chinese–English CLIR.

In Chinese–English CLIR, the accuracy of dictionary-based query translation is limited by two major factors: translation ambiguity and the presence of out-of-vocabulary (OOV) terms. We explore alternative methods for translation disambiguation, and demonstrate new techniques based on a Markov model and the use of web documents as a corpus to provide context for disambiguation. This simple disambiguation technique has proved to be extremely robust and successful. Queries that seek topical information typically contain OOV terms that may not be found in a translation dictionary, leading to inappropriate translations and consequent poor retrieval performance. Our novel OOV term translation method is based on the Chinese authorial practice of including unfamiliar English terms in both languages. It automatically extracts correct translations from the web and can be applied to both Chinese–English and English–Chinese CLIR. Our OOV translation technique does not rely on prior segmentation and is thus free from segmentation error. It leads to a significant improvement in CLIR effectiveness and can also be used to improve Chinese segmentation accuracy.

Good quality translation resources, especially bilingual dictionaries, are valuable resources for effective CLIR. We developed a system to facilitate construction of a large-scale translation lexicon of Chinese–English OOV terms using the web. Experimental results show that this method is reliable and of practical use in query translation. In addition, parallel corpora provide a rich source of translation information. We have also developed a system that uses multiple features to identify parallel texts via a k-nearest-neighbor classifier, to automatically collect high quality parallel Chinese–English corpora from the web. These two automatic web mining systems are highly reliable and easy to deploy. In this research, we provided new ways to acquire linguistic resources using multilingual content on the web. These linguistic resources not only improve the

efficiency and effectiveness of Chinese–English cross-language web retrieval; but also have wider applications than CLIR.

Chapter 1

Introduction

The rapid spread of the web and improvements in general information retrieval (IR) techniques have allowed people worldwide to access vast quantities of information. With the increase of multilingual information available on the web and the growing number of non-native English speakers browsing the Internet, it has become increasingly valuable to have IR systems that can retrieve relevant information regardless of language boundaries. A cross-language information retrieval (CLIR) system retrieves documents in a language that is different from the query language [Oard, 2001]. An example showing the usefulness of such a system is when a user might have some knowledge of the document language but has difficulty in formulating effective queries. These users might very well be able to distinguish relevant documents from irrelevant documents based on their limited knowledge. Such users could then send the documents that they have judged relevant on to a translation service bureau or machine translation system. The advantages of a CLIR system are not limited to individual users of the Internet. Many business, government, social, and multi-national organizations can also benefit from the ability to perform searches across different languages. Groups such as investment companies, for example, collect historical reports and legal documents from around the world for decision making. A patent officer searches foreign patent databases or a security officer monitors foreign news and intelligence sources. Regardless of the task, the development of effective CLIR systems to enable searching between different languages can provide access to the information that is not otherwise accessible to people who have limited knowledge of other languages.

While English is the most widely used language on the web, the use of Chinese as a query language has grown rapidly in recent years, to the point where it is now the third or fourth

most popular query language. Most Chinese web users have limited English vocabulary and thus it can be difficult for them to formulate effective English queries. They would like to retrieve relevant English documents on the web using queries expressed in Chinese, especially in instances where information available in English is more plentiful and detailed than that in Chinese. For example, general information on admissions to PhD programs at foreign universities, overseas tourist information, or foreign business law and regulations. Such information needs have given rise to greater interest in Chinese–English CLIR. CLIR between Western and Asian languages poses significant problems due to the great differences in the structural and written forms of the languages.

The goal of this research is to develop new techniques to improve the performance of Chinese–English CLIR. We concentrate on short queries as they represent typical web queries and have proved to be difficult to translate due to lack of context. Within this area, our focus is on two related topics: first, improved query translation via translation disambiguation and enhanced dictionary coverage; and second, automatic acquisition of translation resources through periodically crawling the web. Additionally, we evaluate a novel use of mutual information to select additional query terms in post-translation query expansion. We also use a series of experiments to identify the value of the individual components of our post-translation expansion process and to explore the sensitivity to parameter settings. We show that post-translation query expansion can be used to improve CLIR effectiveness**,** especially for imperfectly translated queries.

1.1 Improved Query Translation

There are several approaches to the implementation of CLIR. Perhaps the most popular is to use a dictionary to translate the queries into the target language and then use mono-lingual retrieval [Kishida et al., 2004; 2005; Braschler and Peters, 2003; Braschler et al., 2004]. As with other CLIR language pairs, in Chinese–English CLIR the accuracy of dictionary-based query translation is limited by two major factors: the presence of out-of-vocabulary (OOV) terms and translation ambiguity.

OOV term translation

Queries that seek topical information typically contain OOV terms that are not found in a translation dictionary, leading to inappropriate translations and consequent poor retrieval performance. For example, the query may concern current affairs, and thus contain new words or translated words that are outside the scope of the translation dictionary; or the query may contain proper

nouns — such as brand names, place names, or personal names — that are not included in the translation dictionary. Although only some queries contain OOV terms, incorrect translation of such terms almost inevitably leads to disastrous results.

Successful translation of OOV terms is one of the challenges of CLIR. Particular difficulties arise in languages where there are no clearly defined boundaries between words, as is the case with Chinese text. Existing Chinese–English CLIR systems use a range of solutions, such as transliteration techniques [Meng et al., 2004; Lin and Chen, 2002], web mining [Lu et al., 2002; Cheng et al., 2004; Wang et al., 2004], and approaches based on parallel corpora [McEwan et al., 2002; Yang and Li, 2002; Chen and Nie, 2000] to the problem of OOV terms. However, they appear to require manual intervention in order to correctly segment OOV terms.

We demonstrate a novel OOV term translation technique, based on the Chinese authorial practice of including unfamiliar English terms in both languages. By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we show that we are then able to infer an appropriate translation with reasonable confidence. We circumvent segmentation difficulties in OOV translation by using the entire Chinese query to search on the web. The idea of using the web to search for translations is not new [Chen et al., 2000; Meng et al., 2004]; however, our technique can extract translations that were previously undetected, or only detected after manual intervention to provide correct segmentation. After using the entire Chinese query to fetch web documents written in Chinese, we collect the English text that is preceded by any substring of the original Chinese query. In formulating our approach, we also consider English text that is not immediately adjacent to the Chinese query terms. However, such text is found to be only rarely a reliable translation. By applying simple statistical techniques based on frequency and length analysis, we extract the best Chinese–English translations. Even when we were only able to find a small number of Chinese–English co-occurrences, our approach still proved to be robust.

We test our OOV term translation technique on several collections and a set of terms from news articles and show that this technique is robust and provides a substantial improvement in CLIR effectiveness. Our technique can be applied to both Chinese–English and English–Chinese CLIR, correctly extracting translations of OOV terms from the web automatically, and thus is a significant improvement on earlier work. In addition, the extracted Chinese terms can be used to enhance a Chinese segmentation dictionary and thus improve Chinese segmentation accuracy.

Translation disambiguation

Translation ambiguity stems from the fact that many words do not have a unique translation, and sometimes the alternative translations have very different meanings. This problem is particularly severe in view of the observed tendency of web users to enter short queries; in general it is difficult for even a human to reliably determine the intended meaning from the available context. By using simple dictionary translations without addressing the problem of translation ambiguity, the effectiveness of CLIR can be 60% lower than that of monolingual retrieval [Ballesteros and Croft, 1998; Aljlayl and Frieder, 2001]. Researchers have used different disambiguation techniques utilizing statistics obtained from the pre-defined training corpora, all seemingly with good results. These have included using term similarity [Adriani, 2000; Maeda et al., 2000], co-occurrence statistics in the target document collection [Ballesteros and Croft, 1998; Gao et al., 2002], and probabilistic methods based on a language model [Federico and Bertoldi, 2002].

We propose an improved technique for disambiguation. The major differences in our proposal are: first, we integrate the concept of window size effect and distance factor into the language model; second, we use statistics obtained from the web documents as a corpus, rather than a specific test collection, to provide context for disambiguation.

Our translation disambiguation technique is based on a Markov model [Markov, 1971]; such models have been used widely for probabilistic modelling of sequence data. Given a query $s_1, s_2, s_3,\ ...,\ s_n$, each translation candidate set T is a sequence of words $t_1, t_2, t_3,\ ...,\ t_n$. We use a bigram Markov model $P(T) = P(t_1, t_2, t_3,\ ...,\ t_n)$ to estimate the maximum likelihood of each sequence of words, and select the translation set T with the highest $P(T)$ among all possible translation sets. We observe that two words being in close proximity generally provides stronger correlation and produces more credible results for disambiguation of translation than does co-occurrence of two words in a large window. Therefore, we investigate the effects of distance factor and window size when using a Markov model to provide disambiguation.

Techniques that use statistics obtained from training corpora for translation disambiguation, have two obvious drawbacks: first, a large training corpus is not always available and the construction of a large corpus needs many human resources; and second, any particular training corpus will always have limited coverage and will not satisfy the open domain problem. The growth of the web has made available vast written and spoken resources on a global scale from almost all countries in the world. The web can be used as **a** practical resource for estimating the coherence of the translated terms. We assume that the correct translations are generally semantically related and tend to co-occur more often in documents than do incorrect translations. Rather than

collect corpus statistics from the specific collection, we make use of web documents extracted by a search engine as a corpus to disambiguate dictionary translation. In order to retrieve a set of web documents that contain the candidate translations, we demonstrate how structured queries are constructed using Boolean operators, and consequently prevent query terms with large numbers of candidate translations dominating the retrieval process.

We compare the effectiveness of our technique to those of other approaches, based on the *mutual information* measure [Gao et al., 2002] and *Dice similarity coefficient* measure. [Adriani, 2000] using the same data sets (NTCIR 4 and 5). Our experimental results show that although each of the techniques uses different models, formulae, parameters, and translation selection algorithms; nonetheless each achieved comparable results across multiple data sets. Our results also show that, when using the web, it is possible to achieve effectiveness comparable to that obtained with a pre-defined training corpus.

1.2 Automatic Linguistic Resources Acquisition

Linguistic resources such as bilingual lexicons and parallel web text are essential in CLIR. The effectiveness of a CLIR system is inevitably limited by the calibre of translations; clearly resources with wide coverage and extensive information are preferable. However, high quality linguistic resources are typically difficult to obtain and exploit, or expensive to purchase. The amount of multilingual information available on the web is expanding rapidly, and provides a valuable new set of linguistic resources. In this research, we develop two automatic mining systems to make use of publicly available translation resources on the web:

- AutoLex, a system to extract topical translations from Chinese text on the web and automatically update Chinese–English (as well as English–Chinese) translation lexicons via periodic crawls of the web;
- Web Parallel Data Extraction (WPDE), a system to automatically collect high quality Chinese–English parallel corpora from the web.

AutoLex — OOV term lexicon construction system

The problem of translation is particularly acute for topical terms, as unfamiliar English terms are typically absent from static lexicons. If we can automatically discover the translations of new words from web text, we can improve dictionary coverage and alleviate the OOV problem. By periodically revisiting the web to discover new translations, coverage can be continually and automatically

updated. Moreover, such a technique has wider applications than CLIR; for example, discovered translations of OOV terms can guide development of printed translation dictionaries and can be used in machine translation of full text.

We develop AutoLex system to facilitate construction of a large-scale translation lexicon of Chinese–English OOV terms using the web. AutoLex consists of three stages: web site selection, web page processing, and OOV translation extraction. Rather than using a specific corpus to identify OOV terms, we mine the web as broadly as possible. We started by using a set of randomly selected English seed terms to locate Chinese web sites that seemed likely to contain a mixture of Chinese and English text. Once new OOV terms are extracted from these sites, we use them to locate further web sites. We then mine these sites to extract bilingual snippets and combine syntactic structure analysis with co-occurrence statistics to infer translations. An added advantage of this approach is that, by first detecting English OOV terms and then extracting Chinese translations, we avoid problems associated with Chinese segmentation — a particular problem in the presence of OOV terms, since there is no prior segmentation information available. For some English terms, there are many candidate instances in web documents where the term is given in both languages, but the translations and usage are often inconsistent; in these cases, we show that co-occurrence statistics can be used to identify good translations. In other cases, there are only a few examples of translation for a given term; we show that these too can be used, by making use of the structure of Chinese text.

To evaluate our system, we crawl 80 gigabytes of Chinese text and extracted a large number of candidate translations. We measure the effectiveness of the system in two ways: human assessment of the quality of translations extracted, and their usefulness in query translation, based on NTCIR and TREC data. These results, although based on only a tiny fraction (around 0.03%) of the Chinese web, show that a web-based lexicon is reliable and effective.

WPDE — Web parallel data extraction system

Parallel corpora provide a rich source of translation information. They have been used to train statistical translation models [Nie et al., 1999; Franz et al., 2001; Brown et al., 1990], translation disambiguation systems [Ballesteros and Croft, 1998], OOV term translation [McEwan et al., 2002], and multilingual thesaurus construction [Chau and Yeh, 2001]. However, some parallel corpora are subject to subscription or licence fee and thus not freely available, while others are domain specific.

In order to take advantage of publicly available parallel corpora, we develop WPDE system,

that combines multiple features to identify parallel texts via a k-nearest-neighbor (KNN) classifier, to automatically collect high quality parallel Chinese–English corpora from the web. WPDE uses a three stage process: first, candidate sites are selected and crawled; second, candidate pairs of parallel texts are extracted; finally, we validate the parallel text pairs. Compared to previous systems [Nie et al., 1999; Resnik and Smith, 2003; Chen et al., 2004], WPDE contains improvements at each stage. Specifically, in stage one, in addition to anchor text, image ALT text (the text that always provides a short description of the image and is displayed if an image is not shown) is used to improve the recall of candidate sites selection. In stage two, candidate pairs are generated by pattern matching and an edit-distance similarity measure, whereas previous systems only applied one or the other of these. In stage three, where previous systems used a single principle feature to verify parallel pages, WPDE applies a KNN classifier to combine multiple features.

Experiments on a large manually annotated data set show that each of the methods leads to improvements in terms of the overall performance in each step, and that the combined system yields the best overall result reported. Our results also show that the use of the KNN classifier with multiple features achieves substantial improvements over the systems that use any one of these features. WPDE has achieved a precision rate of 95% and a recall rate of 97%, and thus is a significant improvement over earlier work.

1.3 Research Contributions

In this research, we identify the problems confronting the task of dictionary-based query translation in Chinese–English CLIR and develop a set of techniques for translation disambiguation and OOV term translation. A combination of these techniques provides a significant improvement in CLIR effectiveness, and allows us to achieve up to 97% of monolingual retrieval effectiveness. In addition, we developed two automatic web mining systems to extract and build high quality linguistic resources — a lexicon of topical terms and a parallel corpus — for Chinese–English CLIR. These two systems are shown to be robust, effective, and easy to implement.

This research makes the following contributions to the field of Chinese–English CLIR:

- We proposed a new technique based on a Markov model and the utilization of web documents as a corpus to provide context for Chinese–English query translation disambiguation. This simple technique has proved to be extremely robust and successful.

- We developed a new segmentation-free technique to identify Chinese OOV terms and extract English translations. This technique does not rely on prior segmentation and is thus free

from segmentation error. It leads to a significant improvement in CLIR effectiveness; it can also be used to improve Chinese segmentation accuracy.

- Our automatic web mining systems (AutoLex and WPDE) are highly reliable and easy to deploy. In this research, we provided new ways to acquire linguistic resources using multilingual content on the web. These linguistic resources not only improve the efficiency and effectiveness of Chinese–English cross-language web retrieval; but also have wider applications than CLIR.

1.4 Thesis Structure

Chapter 2 reviews approaches to CLIR. This includes different approaches to translation in Chinese–English CLIR, the issues of dictionary-based query translation, and existing approaches to these problems. We also discuss the characteristics of Chinese text and practical solutions to the challenges of Chinese text processing. Additionally, we describe IR system evaluation and our CLIR experimental methodology.

Chapter 3 presents a novel method to dynamically discover translations of OOV terms through the mining of web text. Our method does not rely on prior segmentation and is thus free from segmentation error.

Chapter 4 demonstrates new techniques based on a Markov model and the utilization of web documents as a corpus to provide context for translation disambiguation.

Chapter 5 concerns Chinese–English query translation experiments, including our experimental setup and results. We also present our two-stage procedure of selecting additional query terms for post-translation query expansion based on term weighting and word association information.

Chapter 6 outlines the architecture of the AutoLex system developed to extract topical translations from Chinese text on the web and automatically update Chinese–English (as well as English–Chinese) translation lexicons via periodic crawls of the web.

Chapter 7 presents the WPED system for automatic mining high quality Chinese–English parallel corpora from the web.

Chapter 8 concludes the thesis with a brief description of the contributions of this work and discuss directions for future research.

Chapter 2

Background

The most prevalent language for communication on the web is English. According to a current study [Gey et al., 2005] (as shown in Table 2.1[1]), there are approximately 2.9 billion web pages on the Internet by the year 2003, and nearly 57% of all web pages are written in English. However, the number of languages used on the web is diversifying. Non-English speakers are the fastest-growing group of new web users and there is a growing interest in non-English sites as the web becomes truly multi-lingual (see Figure 2.1[2]). As of September 2004, over 64% of the global web users are non-English speakers, according to "Global Internet Statistics 2004" by Global Reach[3]. In China, for example, the Internet user population increased to more than 104 million by the end of June, 2005 (China Internet Network Information Center Statistical Survey Report[4], published in July 2005). The Global Reach statistics also shows that about 90% of the web users prefer to access the Internet in their native languages.

The language barrier has become a major restriction on global information exchange and knowledge sharing, and its impact is more significant in countries whose language is non-alphabetical, such as China, Japan, and Korea. For example, most Chinese web users have some knowledge of English. Due to limited English vocabulary, they find it difficult to formulate effective English queries. Thus the ability to retrieve relevant English documents using Chinese queries should be considered a necessary part of information access for Chinese users, who desire to identify or monitor developments around the world. Moreover, there may be some instances where the user wants to retrieve documents in foreign languages, especially for instances where information available in

[1] http://ucdata.berkeley.edu:7101/language-distribution-of-the-web-table.htm

[2] http://global-reach.biz/globstats/evol.html

[3] http://global-reach.biz/globstats/index.php3

[4] http://www.cnnic.net.cn/uploadfiles/pdf/2005/7/20/210342.pdf

Language	**Number of web pages**	**Percentage**
English	1,690,901,291	57.37%
German	256,997,640	8.72%
French	120,860,523	4.10%
Russian	98,422,529	3.34%
Japanese	95,372,822	3.24%
Simplified Chinese	93,228,783	3.16%
Spanish	89,429,894	3.03%
Italian	78,104,597	2.65%
Korean	66,390,095	2.25%
Dutch	64,020,603	2.17%
Portuguese	34,193,853	1.16%
Czech	31,429,288	1.07%
Swedish	28,700,267	0.97%
Polish	27,864,012	0.95%
Danish	27,622,742	0.94%
Traditional Chinese	21,421,052	0.73%
Catalan	19,245,884	0.65%
Norwegian	14,728,161	0.50%
Hungarian	14,504,776	0.49%
Finnish	11,856,905	0.40%
Slovak	9,236,475	0.31%
Turkish	7,587,437	0.25%
Greek	5,261,482	0.18%
Hebrew	4,563,942	0.16%
Romanian	3,993,211	0.14%
Arabic	3,882,050	0.13%
Thai	3,547,806	0.12%
Others*	23,959,095	0.93%
Total web pages	2,947,327,215	100.00%

Table 2.1: Language distribution of 2.9 billion web pages in July 2003 (Source: AlltheWeb.com for July 2003, prepared by Takagi and Fredric Gey.) ∗ 22 other languages (in order of count): Croatian, Estonian, Bulgarian, Slovenian, Byelorussian, Icelandic, Lithuanian, Indonesian, Ukrainian, Latvian, Galician, Vietnamese, Malay, Afrikaans, Basque, Latin, Faeroese, Albanian, Frisian, Welsh, Serbian and Swahili each have less than 0.1% *and cumulatively less than one percent of the total web pages.*

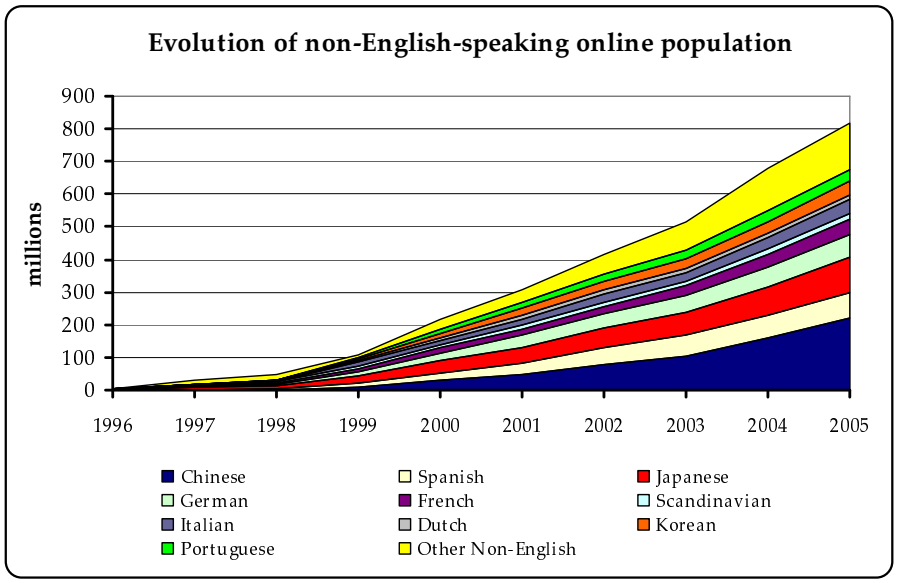


Figure 2.1: Evolution of non-English-speaking online population. (Based on the data from Global Internet Statistics)

languages other than the user's native language is more plentiful and detailed. For example, one would expect trade between China and the U.S. to grow significantly in the near future because of the impending WTO membership for China. Monitoring trends and status information from Chinese sources may be an essential operation for organizations interested in these affairs. The web phenomenon, coupled with increasing globalization of corporations and organizations, has led to strong demand for tools that permit the users to find information regardless of language boundaries. The demand had also been stimulated by more corporations finding themselves competing in a world-wide marketplace driven by foreign language information. These factors have given rise to greater interest in Chinese–English cross-lingual information retrieval (CLIR).

CLIR has the potential to allow people to find documents irrespective of the language used in query or document. It is of growing importance because it can open up a whole world of information for the user, especially with the ease and convenience of access and delivery of foreign documents provided by the web. CLIR has been a research sub-field for more than a decade. The field has sparked three major evaluation efforts: the Text REtrieval Conference (TREC)[5] Cross Language Track, which currently focuses on the Arabic language; the Cross-Language Evaluation Forum (CLEF)[6], a spinoff from TREC covering many European languages; and the NACSIS

[5] `http://trec.nist.gov/`

[6] `http://www.clef-campaign.org/`

Test Collection Information Retrieval (NTCIR)[7] Asian Language Evaluation covering Chinese, Japanese and Korean. These forums provide means for exchanging research results, collections and for cross-lingual system evaluation. The U.S. National Institute of Standards and Technology (NIST) organizes an evaluation of CLIR technology in conjunction with TREC. TREC sponsored the first large-scale evaluations of the retrieval of non-English documents, retrieval of recordings of speech, and retrieval across multiple languages. In previous years, the TREC CLIR evaluations have included retrieval of English, French, German, Italian, and/or Chinese news stories based on queries expressed in another language. The last cross-lingual track in TREC was run in TREC 2002 [Voorhees, 2002a], which focused on the retrieval of Arabic language newswire documents using English topics. The CLIR tasks are studied in both CLEF and the NTCIR workshops. CLEF aims at promoting research and development in CLIR by providing an infrastructure for the testing and evaluation of IR systems operating on European languages, and creating test-suites of reusable data which can be employed by system developers for benchmarking purposes. These objectives are being achieved through the organization of a series of system evaluation workshops. The NTCIR workshop is a forum for evaluating information retrieval techniques for Asian languages. It is designed to enhance the research in information access technologies such as information retrieval, question answering, text summarization, text mining, by providing infrastructure of evaluation and testing including large-scale reusable test collections, a forum of researchers, investigation of evaluation methodologies and metrics. The workshop is held once per one and half years. The NTCIR workshop started in late 1998 with emphasis on Japanese and English, then gradually enlarged the scope to East Asian languages, i.e. Chinese and Korea, and English documents published in East Asia, and has attracted international participation not only from Asia but also from North America, Europe, and Oceania.

We begin this chapter with an overview of translation approaches to CLIR in Section 2.1. The comprehensive literature reviews presented by Oard and Diekema [1998], Peters and Sheridan [2001], and Kishida [2005] described various techniques and methods for enhancing effectiveness of CLIR. Query translation has proved to be the predominate translation approach in current research. In Section 2.2, we therefore describe different ways to perform query translation in CLIR by employing machine techniques, using parallel corpora, or using a bilingual dictionary. Due to the simplicity and the increasing availability of machine readable dictionaries, much of the research effort in CLIR has been put into the dictionary-based query translation, in which queries are translated word-by-word via dictionary lookup. We explore the issues of dictionary-

[7] http://research.nii.ac.jp/ntcir/workshop/

based query translation and review existing approaches to these problems in Section 2.2.3. In terms of text retrieval, there are some inherent linguistic difficulties in Chinese. In Section 2.3, we discuss the characteristics of Chinese text, issues associated with Chinese text retrieval, and practical solutions to the challenges of Chinese text processing. We also describe the Chinese–English bilingual translation dictionaries we have used in our experiments. Section 2.4 describes the effectiveness evaluation of IR systems, followed by a discussions of our CLIR experimental methodology.

2.1 Translation Approaches to CLIR

CLIR is the task of issuing a query in one language and retrieving relevant documents in another language. It aims to benefit the user in finding and assessing information without being limited by linguistic barriers. The language barrier can be bridged by translating the query, translating the documents, or translating both into an intermediate representation that can be either a pre-defined controlled vocabulary or automatically extracted semantic structures that have been extracted from parallel document collections. The different approaches to CLIR are described below.

2.1.1 Direct translation

Within the direct translation framework, two main approaches to CLIR are query translation and document translation. In *query translation* the query is translated into the documents' language, allowing querying via monolingual retrieval [Ballesteros and Croft, 1997]. In *document translation*, the documents are translated into the supported query language, again allowing querying via monolingual retrieval [Oard, 1998].

Intuitively, it seems that document translation would be superior to query translation. Documents provide more context for resolving ambiguities and the translation of source documents into all the languages supported by the IR system effectively reduces CLIR to a monolingual IR task. Furthermore, it has the added advantage that document content is accessible to users in their native tongue. Document translation is inherently slower than query translation but it can be done off-line and translation speed may therefore not be crucial. Document translations need to be stored for indexing though, and storage space may be a limiting factor, especially if many languages are involved. Query translation on the other hand can be improved by consulting the user during translation, an option that is not available for document translation. In a comparison of both query and document translation, Oard [1998] found that document translation generally outperforms query translation because translated documents provide greater linguistic context,

which in turn facilitates part-of-speech disambiguation and sense selection.

Several integrated translation methods have also been proposed for CLIR. McCarley [1999] showed that a hybrid system, where the relevance degree of each document is the mean of those obtained with query and document translation methods, outperformed systems based on either query or document translation methods. Fujii and Ishikawa [2000] proposed a two-stage method, which also integrates the query and document translation methods. In the first stage, the query translation method is used to retrieve a limited number of foreign documents. Then, in the second stage, retrieved documents are machine translated into the user language. Thus, the computational cost required for the MT-based document translation can be minimized. Finally, those documents are re-ranked based on the score, combining those individually obtained with the first and second stages. Preliminary experiments using the NTCIR (1 and 2) Japanese–English CLIR collections showed that the two-stage method outperformed the query translation method. Chen and Gey [2004] showed that translating the entire document collection outperforms query translation and also that a combination of query translation and document translation can lead to further improvements in retrieval effectiveness.

Nevertheless, query translation is the prevailing CLIR method in current research, because document translation is time-consuming and requires re-indexing of the entire collection each time the automatic translation system is modified to produce a new target collection. Query translation using machine translation systems or machine readable dictionaries has been adopted by almost all groups participating in the NTCIR (4 and 5) CLIR task [Kishida et al., 2004; 2005], and proved to be the predominant translation approach in CLEF (2003 and 2004)[Braschler and Peters, 2003; Braschler et al., 2004].

2.1.2 Transitive translation

Sometimes, there is no direct translation resource between the query language and the document language. For example, there are not always good dictionaries even between common European languages. Direct translation from language A into language B may therefore not be possible. However, there might be a dictionary between language A and language C, and one between language C and language B, which means that translation would be possible first from A into C and then from C into B. This kind of translation through an intermediate (or pivot) language is called *transitive translation*. It may reduce the number of translations needed when translations have to be performed between a large number of languages. If there are 50 languages and a translation system is needed between each pair of them, then 2450 translation systems will be

required. If there is a single pivot language that can be used in all the translations, only 98 translations are needed (49 translations from the source languages into the pivot language, and 49 from the pivot language into the target languages).

Unfortunately, transitive translation is prone to errors due to the ambiguity introduced by double translation. In previous studies, transitive translation has been much less effective than direct translation. Some techniques have been used to try to improve the performance of transitive translation. Gollins and Sanderson [2001] tried to solve the problem of ambiguity in transitive translation through triangulation, for example, by using several translation routes. They used several pivot languages and merged the translation results from the different routes. Triangulation can have two positive effects. First, it can have a query expansion effect, introducing different new terms by its different pathways. Secondly, it can eliminate some translation ambiguity found within a single pathway by requiring that the translation of a term must be found via more than one translation pathway. The first benefit is best achieved by taking the union of the transitive translation sets as the final query; the second benefit is best achieved by using the intersection of those sets. This indeed had a favorable effect. On the whole, the effectiveness was low, mainly because of the poor translation resources used. Ballesteros [2000] reduced the ambiguity of transitive translation by query structuring and various expansion techniques. Using simple word-by-word translation from Spanish into French via English, the transitive translation achieved an average precision 91% below the baseline direct translation. Adding query structure to transitive translations yielded significant improvements in retrieval effectiveness. Using this method, effectiveness was improved to 34% below the direct translation.

2.2 Strategies for Query Translation

Recent CLIR research [Gao et al., 2001; Federico and Bertoldi, 2002; Kraaij et al., 2003; Cheng et al., 2004; Monz and Dorr, 2005] has concentrated on query translation. There are three main strategies for CLIR query translation: machine translation, a parallel corpus, or a bilingual dictionary.

2.2.1 Machine translation

Machine Translation (MT) based query translation method uses existing machine translation techniques to provide automatic translations. Using MT systems for query translation is also quite popular when such a system is available for the particular language pairs. In the Eighth Text REtrieval Conference evaluations for CLIR, at least half of all groups used the SYSTRAN

machine translation system in some form for parts of their experiments [Braschler et al., 1999]. The reason is that Systran covers all four languages (English, German, Spanish, and Italian) included in TREC, and it is easily accessed through the web.

However, empirically, the effectiveness of MT-based approaches so far have been inconclusive. Oard [1998]'s TREC−6 experiments suggest that the effectiveness may depend on the types of the queries. For short queries with one to three words, his results with LOGOS, a commercial MT system, were worse than those of dictionary-based query translation (discussed below); for long queries, consisting of a few sentences, MT-based document translation produced better results than MT-based query translation, which was in turn better than dictionary-based query translation. The TREC−7 experiments by Nie [1998] partly supported Oard's observation: translating sentence-based queries, when using SYSTRAN, Nie obtained better results than those obtained by using corpus-based and dictionary-based methods. However, he did not report parallel experiments with short queries for comparison. The TREC results of Gey and Jiang [1999] from UC Berkeley are even more interesting. They found that SYSTRAN outperformed dictionary lookup on the TREC−7 CLIR corpus, containing news stories from the *Associated Press* and the *Swiss News Agency*, but significantly under-performed dictionary lookup (0.1063 versus 0.2707 in average precision) on the *GIRT* document collection available in TREC−8 in the field of social science when using a dictionary automatically extracted from an existing bilingual thesaurus in the same field. It is also worth mentioning that the development cost of rule-based MT system is typically a few person-decades per language pair, and the commercially available MT systems only exist for a few language pairs, typically the most common languages. Therefore, MT is not regarded as a promising method for query translation in CLIR.

2.2.2 Parallel corpus

A parallel corpus consists of a collection of texts that have been translated into one or more other languages. Over the past 10 to 15 years, the increased availability of computational power, memory, storage, and parallel texts has enabled active research in the field of corpus-based translation — using manually-translated texts as the basis for translating new texts. Corpus-based machine translation can be subdivided into three categories: translation memories, example-based machine translation, and statistical machine translation.

Translation memories simply store prior reference translations of text and retrieve the nearest match in their database. It is then up to the human translator to clean up the retrieved translation and to account for the differences between the entry in the database and the actual passage to be

translated. This very simple technology has resulted in a major increase in human translator productivity, and is commercially available in several products, including IBM's Translation Manager (TM2), STAR's Transit, and SDL International's SDLX.

Example-based machine translation approaches are often characterized by their use of a bilingual corpus as their main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning [Brown, 1996]. This method is grounded on the conviction that there are no pre-established solutions to translation, but most possible solutions can be found in texts already translated by professionals. In other words, a large portion of a translator's competence is encoded in the language equivalences that can be found in already translated texts.

Statistical machine translation systems use a parallel training corpus to build a probabilistic model of the possible translations for words and the reordering of words between languages. After training, the corpus is no longer required, as all translation is performed using the statistical model. Approaches that involve training a statistical translation model have been explored by, for example, Nie et al. [1999] and Franz et al. [2001]. In Nie et al.'s approach, statistical translation models (usually IBM model 1) are trained on a parallel corpus. The models are used in a straightforward way: the source query is submitted to the translation model, which proposes a set of translation equivalents, together with their probabilities. The latter are then used as a query for the retrieval process, which is based on a vector space model. Franz et al.'s approach uses a better founded theoretical framework: the OKAPI probabilistic IR model. The present study uses a different probabilistic IR model, one based on statistical language models [Hiemstra et al., 2001; Xu et al., 2001]. This IR model facilitates a tighter integration of translation and retrieval. An important difference between statistical translation approaches and approaches based on document alignment discussed above is that translation models use alignment at a much more refined level. Consequently, the alignments can be used to estimate translation relations in a reliable way. On the other hand, the advantage of the CLIR approaches that rely only on alignment at the document level is that they can also handle comparable corpora, that is, documents that discuss the same topic but are not necessarily translations of each other.

Most previous work on parallel texts has used a few manually constructed parallel corpora, notably the *Canadian Hansard* corpus. This corpus contains many years' debates in the Canadian parliament in both English and French, amounting to several dozens of millions of words in each language. Documents from the European parliament represent another large parallel corpus in several European languages. However, the availability of this corpus is much more restricted than

the Canadian Hansard. The Hong Kong government publishes official documents in both Chinese and English. They form a Chinese–English parallel corpus, but again, its size is much smaller than that of the Canadian Hansard. For many other languages, no large parallel corpora are available for the training of statistical models. The explosion of information on the web yields an interesting new source of parallel text, which researchers have been harnessing through automated retrieval from the web. Finding a realistic, systematic, and cost-effective way for automated acquisition of bilingual parallel corpora remains an open challenge in corpus-based CLIR.

Kraaij et al. [2003] developed the PTMiner to mine large parallel corpora from the web. PTMiner used search engines to pinpoint candidate sites that are likely to contain parallel pages, and then used the URLs collected as seeds to further crawl each web site for more URLs. The pairs of web pages were extracted on the basis of manually defined URL pattern-matching, and further filtered according to criteria such as file length, HTML structure, and language character set. Resnik and Smith [2003] at the University of Maryland reported a similar effort. Their approach is to use a web spider to collect pages containing certain key expressions that tend to indicate that a certain link points at a translated version of the page, and then filter the retrieved pairs of pages by ensuring structural parallelism of the HTML tags within the pages. However, it is usually not only time consuming but also expensive to acquire large high-quality parallel bilingual corpora, particularly for minor languages.

2.2.3 Dictionary based methods

Due to the increasing availability of machine readable dictionaries, much of the research effort in CLIR has been put into the dictionary-based approaches, in which queries are translated by looking up terms in a bilingual dictionary and using some or all of the translated terms. It is often easier to use dictionary-based translation approaches for query translation rather than parallel corpus based methods. However, the performance of dictionary-based approaches are limited by three factors: phrases and compound words, translation ambiguity, and out-of-vocabulary terms.

Phrases and compound words

Failure to translate multi-term phrases has been shown to be one of the factors responsible for the errors associated with the dictionary-based method [Ballesteros and Croft, 1997]. Hull and Grefenstette [1996] also showed that the performance achieved by manually translating phrases in queries is significantly better than that of a word-by-word translation using a dictionary. A technique that can be used to alleviated the impact of the phrases and compound words translation

is by identifying phrases in the query and translating them using a phrase look-up dictionary. However, language use is a creative activity. New phrases are continually created. It is unrealistic to expect a "complete" phrase dictionary. A critical problem remains: if a phrase is not stored in a lexicon, how can one identify it in a query and translate it correctly?

Ballesteros and Croft [1998] identified phrases as sequence of nouns and adjective-noun pairs using a set of syntactic patterns; and then performed phrase translation using information on phrase and word usage in the *Collins* Spanish–English machine readable dictionary (MRD). This allowed the replacement of a source phrase with its multi-term representation in the target language. When a phrase could not be identified using this information, the remaining phrase terms were translated in one of two ways. Terms were either translated word-by-word using parallel corpus disambiguation, or they were translated as multi-term concepts using the co-occurrence statistics. The parallel corpus disambiguation method uses query context to disambiguate all remaining terms whether or not they are constituents of a phrase. The co-occurrence method disambiguates the remaining phrase term translations based on their co-occurrence with other terms in a phrase. They demonstrated that certain types of noun phrases are easily translated via MRD. However, dictionaries do not provide enough context for accurate phrasal translation in other cases. The correct translations of phrase terms tend to co-occur and incorrect translations tend not to co-occur. Corpus analysis can exploit this information to significantly reduce the ambiguity of phrasal translations. Combining phrase translation via phrase dictionary and co-occurrence disambiguation brings CLIR performance up to 79% of that of monolingual IR.

Gao et al. [2001] proposed an approach to identify noun phrases in English queries using a unified statistical model [Xun et al., 2000], and then translate the identified phrases using a set of translation patterns and probabilities of translated phrases obtained from a language model. The remaining query terms are translated word-by-word. Their statistical model also makes use of lexical information and every query term is taken into account during noun phrase identification. Finally, the Viterbi algorithm [Church, 1988] is applied to make global search in the English query at the sentence level, and thus to obtain linear complexity for the entire process.

Translation ambiguity

By using simple dictionary translations without addressing the problem of translation ambiguity, the effectiveness of CLIR can be 60% lower than that of monolingual retrieval [Ballesteros and Croft, 1998; Aljlayl and Frieder, 2001]. Translation ambiguity stems from the fact that many words do not have a unique translation, and sometimes the alternate translations have very differ-

ent meanings. This problem is particularly severe in view of the observed tendency of web users to enter short queries; in general it is not possible for even a human to determine the intended meaning from the available context. Web search engine log analysis have revealed, that the average query length for a web search was about 2.4 words in English [Spink et al., 2001] and 3.18 characters in Chinese [Pu et al., 2002]. The dictionary-based approaches are prone to errors due to the high possibly of selecting the wrong translation of a term from among the translations provided by the dictionary. Various techniques have been proposed to reduce the ambiguity and errors introduced during query translation. Instead of using all possible translations in the machine readable dictionary, researchers have proposed techniques based on term co-occurrence [Ballesteros and Croft, 1998; Gao et al., 2002], term similarity [Adriani, 2000; Maeda et al., 2000], and language modelling [Federico and Bertoldi, 2002]. A different approach to reducing the effect of the ambiguity problem is to combine the results of translating query terms using a general and a domain-specific dictionary, and then using structural tags to indicate the contextual relationship among the resulting terms [Pirkola, 1998].

Term co-occurrence. Ballesteros and Croft [1998] described a technique that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary translation. Their hypothesis is that the correct translation of query terms should co-occur in target language documents and incorrect translation should tend not to co-occur. They measure the importance of co-occurrence of the elements in a set by the *em* metric. This is a variation of the Expected Mutual Information Measure (EMIM) and measures the percentage of the occurrences of term a and term b that are net co-occurrences (co-occurrences minus expected co-occurrences):

$$em(a,b) = \max\left(\frac{n_{ab} - n_a \times n_b \;/\; N}{n_a + n_b}, 0\right)$$

where

n_a: the number of occurrences of a in the collection

n_b: the number of occurrences of b in the collection

n_{ab}: the number of times both a and b fall in a text window size of t

N: the number of text windows in the collection

Each set is ranked by the *em* score and the highest ranking set is taken as the appropriate translation. Ballesteros and Croft [1998] demonstrated the effectiveness of translating phrases in Spanish queries into English phrases using terms which co-occur in the English collections. Compared to the corpus-based methods previously described in Section 2.2.2, co-occurrence appears to be

significantly better at disambiguation. Building on the work of Ballesteros, Gao et al. [2002] observed that the correlation between two terms is stronger when the distance between them is shorter. They extended the previous co-occurrence model by incorporating a distance factor $D(x, y) = e^{-\alpha(\mathrm{Dis}(x,y)-1)}$. This factor decreases exponentially when the distance between two terms x and y increases, where α is the decay rate, which is determined empirically, and $\mathrm{Dis}(x, y)$ is the average distance between x and y in the collection. They experimented on the TREC−9 Chinese collection and showed that this improved co-occurrence model performs generally better than the basic co-occurrence model. They were able to achieve 84% of the monolingual retrieval effectiveness when using $\alpha = 0.8$ in this decaying co-occurrence model.

Term similarity. Adriani [2000] proposed a translation disambiguation technique based on the concept of statistical term similarity for selecting the best Indonesian translation of an English term from all possible translations given by a bilingual dictionary. These techniques use a term-similarity matrix built using the statistical term-distribution parameters obtained from the Indonesian corpus for Indonesian terms, and a subset of their English collections for English terms. The degree of similarity or association-relation between terms is obtained using a term association measure, the *Dice similarity coefficient* [Rijsbergen, 1979], which is commonly used in document and term clustering. The Dice coefficient is a term based similarity measure $(0 - 1)$ whereby the similarity measure is defined as twice the number of terms common to compared entities divided by the total number of terms in both tested entities. The coefficient result of 1 indicates identical vectors as where a 0 equals orthogonal vectors. The term association value between term x and y, SIM_{xy}, is calculated using the formula below:

$$\mathrm{SIM}_{xy} = 2\sum_{i=1}^{n}(w'_{xi} \times w'_{yi}) \Big/ \left(\sum_{i=1}^{n} w_{xi}^2 + \sum_{i=1}^{n} w_{yi}^2\right)$$

where

w_{xi} = the weight of term x in document i
w_{yi} = the weight of term y in document i
w'_{xi} = w_{xi} if document i also contains term y, or 0 otherwise
w'_{yi} = w_{yi} if document i also contains term x, or 0 otherwise
n = the number of documents in the collection

The disambiguation algorithm computes the sum of maximum similarity values between each candidate translation of a query term and the translations of other terms in the query. For each Indonesian query term, the English translation that has the highest sum is chosen as the

query term's translation. Their term similarity techniques bring Indonesian–English and English–Indonesian CLIR effectiveness to 58% and 74% of monolingual retrieval, respectively. Maeda et al. [2000], working on Japanese–English CLIR, have used a search engine to collect the co-occurrence information between terms in web documents, and applied a modified Dice coefficient to calculate the mutual information between terms. They used one document as the window of co-occurrence.

Language modelling. Federico and Bertoldi [2002] presented a novel statistical model for dictionary-based query translation. Translations are selected using a hidden Markov model (HMM) [Rabiner, 1990], which couples a translation lexicon with a bigram language model (LM) in the target language. Their query-translation model computes the probability of any query-translation pair. This probability is modelled by a HMM in which the observable variable is the Italian query i, and the hidden variable is its English translation e. According to the HMM, the joint probability of a pair (i, e) is decomposed as follows:

$$Pr(i = i_1, ..., \ i_n, \ e = e_1, ..., \ en) = Pr(e_1)Pr(i_1|e_1) \prod_{k=2}^{n} Pr(e_k|e_{k-1})Pr(i_k|e_k)$$

This formula considers two different conditional probabilities: the term translation probabilities $Pr(i|e)$ and the target LM probabilities $Pr(e|e')$. Probabilities $Pr(i|e)$ are estimated from a bilingual dictionary as follows:

$$Pr(i|e) = \frac{\delta(i, e)}{\sum_{i'} \delta(i', e)}$$

where $\delta(i, e) = 1$ if the English term e is one of the translations of Italian term i and $\delta(i, e) = 0$ otherwise. Probabilities $Pr(e|e')$ are estimated on the target document collection, through an order-free bigram (bag-of-word) LM, which tries to compensate for different word positions induced by the source and target languages. Let

$$Pr(e|e') = \frac{Pr(e, e')}{\sum_{e''} Pr(e'', e)}$$

where $Pr(e|e')$ is the probability of e co-occurring with e', regardless of the order, within a text window of fixed size. Smoothing of this probability is performed through absolute discounting and interpolation as follows:

$$Pr(e|e') = \max \left\{ \frac{C(e, e') - \beta}{N}, 0 \right\} + \beta Pr(e)Pr(e')$$

where $C(e, e')$ is the number of co-occurrences in the collection, $Pr(e)$ is the word probability of e over the collection. The absolute discounting term β is equal to the estimate proposed by Ney et al. [1994]:

$$\beta = \frac{n_1}{n_1 + 2 \times n_2}$$

with n_k representing the number of term pairs occurring exactly k times within the collection. Federico and Bertoldi [2002] successfully used the above formula to compute the probability of term e' and e within a fixed text window through an order-free bigram LM in their work. Their experimental evaluation of the CLIR model was performed on the Italian–English bilingual track data used in the CLEF 2000 and CLEF 2001 evaluations. Their experimental results using the 1-best, 5-best, and 10-best translations indicate that on average, higher relative improvements are observed with short topics especially with the 1-best modality; with long topics the 5-best modality provides the highest improvements; using more than one translation slightly improves performance just for long topics. Their best mean average precision value was about 80% of monolingual.

Query structure. Pirkola [1998] explored the effects of query structure and various setups of translation dictionaries on the performance of CLIR. Their study used a structuring method in which English translations that were derived from each Finnish query term were grouped into a set. The derived translations in the same set were combined by the *syn-operator*, which treats its operand translations as instances of the same query term. One possible way by which structuring disambiguates CLIR queries is that it enforces conjunctive relationships between query terms. Often those query terms that have only one or two translations are the most important words of a query and, those query terms that have many translations are unimportant words. Thus, when a query is not structured, all translations derived from all query terms are passed into the IR system at the same time; those unimportant query terms and irrelevant translations tend to dominate the depress the effect of important terms. As a result, many non-relevant documents are highly ranked. When translations are grouped into the set, one translation is picked from each set each time and passed into the IR system; thus, important query terms get relatively more weight. A combination of a general MRD and a domain specific medical MRD was used in the Finnish–English query translation. The positive effects of the medical dictionary were mainly due to two factors. First, it contained query terms that were not found in the general MRD. Typically, domain specific terms are prevalent in special dictionaries, whereas in general MRD they are not common. Therefore general MRDs only rarely translate specific terms. Second, it disambiguated word translations. General MRDs often give many equivalents to a query term, whereas special dictionaries usually give only one or two equivalents. The translations provided by special dictionaries are often unambiguous. They found that it is possible to solve successfully the translation ambiguities and polysemy problems if queries are structured and if both general

terminology and domain specific terminology are available in translation. Their test collection consisted of the *Associated Press newswire*, *Federal Register*, and *Department of Energy abstracts* subsets of the TREC collection. This collection contained $514,825$ documents. They used the health related test queries selected from the TREC topics $1-300$ and were able to achieve 77% of monolingual retrieval effectiveness.

Out-of-vocabulary terms

The out-of-vocabulary (OOV) problem arises from the fact that some query terms are not found in translation resources such as bilingual dictionaries and parallel corpora. For example, the query may concern current affairs, and thus contain new words or translated words that are outside the scope of the translation dictionary; or it may contain proper nouns such as brand names, place names, or personal names that are not included in the translation dictionary. Existing CLIR systems use a range of solutions to the problem of OOV terms.

Transliteration. Transliteration is the process that converts an original term in the source language into an approximate phonetic equivalent in the target language. In Chinese, each character represents a syllable. Chinese has relatively few distinct phonemes and many sounds in English are not present in Chinese (and vice versa), so Chinese terms transliterated from English often do not closely resemble the original English pronunciation. Since many English phonemes often map to one Chinese phoneme, backward transliteration — recovery of the English term from the Chinese — is difficult. An additional problem with this process is that a given English term may have more than one Chinese transliterated equivalent. For example, "Michael Jordan" is transliterated into "麥可喬丹" in Taiwan, "迈克尔·乔丹" in the Mainland of China, and "米高佐敦" and "邁爾克·喬丹" in Hong Kong. This is because different Chinese communities transliterate English terms in different ways.

Meng et al. [2004] presented a learning algorithm for transliteration of OOV names from English to Chinese in the context of cross-language spoken document retrieval. They first used a set of hand-crafted phonological rules by adding or dropping proper phonemes to normalize English syllables into consonant-vowel format. This aimed to overcome some of the phonological differences between the two languages. The process of cross-lingual phonetic mapping (CLPM) then applied a set of automatically generated one-to-one phonetic transformation rules to map English phonemes to their Chinese counterparts. These rules were learned from aligned parallel data using *transformation-based error-driven learning* [Brill, 1995]. The pin-yin syllabic constraints were then

added to the Chinese phoneme sequences generated by CLPM for eliminating the errors in the sequences. A phoneme lattice of pin-yin sub-syllables were generated based on a confusion matrix obtained from the mapping differences between reference phonemes and output phonemes. They searched the phoneme lattice exhaustively for Chinese phonetic sequences that could constitute legitimate syllables to create a syllable graph. Finally, a syllable bigram language model was applied together with the probabilities derived from the confusion matrix to search the graph to find the most probable syllable sequence. Their system appears successful where transliteration rules have been strictly applied, but is not always able to pick the best transliteration when several are in use.

Lin and Chen [2002] developed a backward OOV transliteration process that attempts to recover the original English term from the transliterated Chinese term. This approach requires a candidate set of English terms. Chinese terms and the candidate English terms were converted into a common form using the International Phonetic Alphabet, and a similarity function was then applied to select the closest match. Although there are many sounds in English that have no equivalent Chinese sound and vice versa, they exploited the degree of consistency used in the transliteration process. Their system is trained on sample data to learn phonetic similarities, then produces a rank list of possible English name equivalents. Lin and Chen report that the average rank of correct English term was 2.04. In their experiments they usually had the correct transliteration available to them.

When OOV term translation is based on meaning rather than sound, transliteration techniques fail. For example, "胚胎乾細胞" (embryonic stem cell) and "國際太空站" (international space station) are semantic translations and cannot be connected using transliteration. Also, backward transliteration of Japanese and Korean names will generally fail, such as "黑澤明" (Akira Kurosawa), as they have been transliterated via different rules. In this case, schemes such as web mining and approaches based on parallel corpora can be applied.

Anchor text. Lu et al. [2002] exploited the existence of web pages written in different languages that had anchor text pointing to the same page to identify candidate translations. By applying statistical techniques, the top-ranked translation proved to be correct in 53% of cases. While this technique is useful, the key drawback is that it requires a web page relating to the Chinese OOV term, and sufficient interest to cause linking from a foreign language site. Their technique found several company names, but did not appear to find names of individuals, place names, and other such terms that are rarely the subject of a web page.

Parallel documents on the web. McEwan et al. [2002], Yang and Li [2002], and Chen and Nie [2000] attempted to locate parallel documents on the web, and used these to build bilingual dictionaries. However, such approaches suffer from lack of sufficient high-quality parallel texts and limited domain specific vocabulary. Yang and Li [2002] successfully mined parallel Chinese–English documents from the web, but, as is common with parallel mining, considered only a small domain — press releases from the Hong Kong Government. Chen and Nie [2000] also obtained good results in alignment of English–Chinese documents, but only 427 documents from the Hong Kong government were used in their experiments. Huang et al. [2003] presented an automatic approach to extract Hindi–English named entity pairs from a parallel corpus from the *India Today* collection. Their approach adapted and iteratively updated a Chinese–English transliteration model to Hindi–English named entity extraction. For each English named entity in each sentence pair, this approach searches for a Hindi equivalent with minimum transliteration cost and constructs an Hindi–English named entity list from the bilingual corpus. They extract 1000 Hindi–English named entity pairs with a precision of 91.8%. However, the problem associated with this approach is that of availability of parallel corpora.

Web search engines. When an English term is used in Chinese web documents, and there is no generally recognized Chinese translation, different communities and users translate it in different ways. Thus newly translated English terms tend to be accompanied by the original English, typically immediately after the Chinese text. For example, the text might contain "... 墨尔本皇家理工大学（RMIT University）成立于1887 年 ..." where "墨尔本皇家理工大学" is a sequence of Chinese characters and "RMIT University" is the original English term for "墨尔本皇家理工大学".

Several CLIR researchers have used the web as a resource for OOV translation. Cheng et al. [2004] developed the LiveTrans system, which uses anchor text and search results to extract translations of unknown query terms. Each query is submitted to a web search engine to collect the top search result pages, and then co-occurrence and context information between queries and translation candidates is used to estimate their semantic similarity and determine the most likely translations. Wang et al. [2004] proposed a technique for identifying OOV terms within a Chinese corpus prior to query time, and then uses web search engines to determine translations for unknown terms by mining bilingual search-result pages obtained. This approach can enhance a domain-specific bilingual lexicon.

2.3 Chinese Text Processing

The characteristics of the Chinese language are different to those of western languages. In contrast to English, Chinese text is composed of a sequence of non-spaced idiographic characters rather than alphabetical words, and the number of commonly-used Chinese characters exceeds $10,000$. In terms of text retrieval, there are some inherent linguistic difficulties in Chinese, in particular word segmentation and identification unknown word . Since written Chinese texts lack explicit delimiters between words to indicate boundaries, word segmentation is a prerequisite and a major barrier for Chinese text retrieval. Because a Chinese sentence can often be segmented into many different possible word combinations and it is difficult to decide the correct combination. In addition, most Chinese proper nouns, such as names and locations, which are usually the keywords in queries but are often excluded from dictionaries, are difficult to identify. We discuss Chinese word segmentation in Section 2.3.3.

In this section, we discuss the characteristics of Chinese text, issues associated with Chinese text retrieval, and practical solutions to the challenges of Chinese text processing.

2.3.1 Chinese character encoding

Every Chinese character is represented by a two byte code. Several character encoding standards have been established, for example, Unicode, GB, GBK, BIG5, and CNS. Simplified Chinese (GB) and traditional Chinese (BIG5) are the two best-known standards in Chinese-speaking communities. GB characters are the Chinese characters officially simplified by the government of the People's Republic of China (PRC) in an attempt to promote literacy. GB has about $7,000$ simplified Chinese characters and is used for most Chinese-language printing in PRC and Singapore, whereas BIG5 has about $13,000$ traditional Chinese characters and is used in Hong Kong, Macao, Taiwan, and overseas Chinese communities. GB is gradually gaining popularity among many overseas Chinese communities.

A character set is different from a font that supports that character set. While there is some overlap between GB and BIG5, there are also many simplified characters in GB that are not in BIG5, and many traditional characters in BIG5 that are not in GB. Consequently, conversion from GB to BIG5 is not trivial, since many simplified characters map to multiple BIG5 traditional equivalents. Going from BIG5 to GB is easier, since the conversion from traditional to simplified is much less ambiguous.

Because Chinese is written without spaces between words, word segmentation is a particular important issue for Chinese language processing. We discuss Chinese word segmentation in

```
1    愛滋病 [ai4 zi1 bing4] /AIDS (acquired immune deficiency syndrome)/
2    艾滋病 [ai4 zi1 bing4] /AIDS/
3    談判 [tan2 pan4] /to negotiate/negotiation/conference/
4    人權 [ren2 quan2] /human rights/
5    國立 [guo2 li4] /national/state-run/public/
6    人壽保險 [ren2 shou4 bao3 xian3] /life insurance/
7    國際貿易 [guo2 ji4 mao4 yi4] /international trade/
8    海灣戰爭 [hai3 wan1 zhan4 zheng1] /(Persian) Gulf War/
9    海灣 [hai3 wan1] /bay/gulf [body of water]/
```

Figure 2.2: An example of the CEDICT BIG5–English wordlist.

Section 2.3.3.

2.3.2 Properties of Chinese–English dictionaries

Chinese queries in BIG5 are translated into English using two machine-readable dictionaries: the CEDICT BIG5–English dictionary[8] and the GB–English wordlist from the Linguistic Data Consortium[9] (LDC). Chinese query terms that were not found in the CEDICT BIG5–English dictionary were converted from BIG5 to GB encoding using the freely available *hc3* program[10] of conversion for BIG5 and GB codes, and then searched in the LDC GB–English wordlist.

Although there are many web sites offering English–Chinese translation facilities, actual bilingual dictionaries that are machine-readable for program access are rare. The CEDICT project, started by Paul Denisowski in 1997, aims to create an online, downloadable public-domain Chinese–English dictionary. Although it started as a one-person project, contributions from the Internet community have become the major source of new entries. The latest CEDICT BIG5–English dictionary contains 34,553 entries. As shown in Figure 2.2, each Chinese term is followed by its Pinyin in square brackets and English equivalents separated by slashes. Pinyin is the romanisation of the Chinese "written sound". Romanisation approximates Mandarin pronunciation with Western spellings and includes a tone mark to signify the pitch of a word. Pinyin is used by most modern Chinese dictionaries to denote pronunciation of characters. It is also an efficient input method in Chinese computer software.

[8] `http://www.mandarintools.com/cedict.html`

[9] `http://www.ldc.upenn.edu/`

[10] `ftp://ftp.ifcss.org/pub/software/unix/convert/hc-30.tar.gz`

1	专家	/expert/specialist/dab/whiz/whizz/
2	意见	/view/opinion/suggestions/
3	国际货币基金组织	/International Monetary Fund/IMF/
4	亚洲国家	/Asian country/Asian nation/
5	诺贝尔和平奖	/Nobel peace prize/
6	金大中	/Kim Dae Jung (newly elected president of Korea)/
7	炭疽病	/anthrax/anthracic/
8	菌	/germ/bacteria/mold/mushroom/
9	综合症	/syndrome/

Figure 2.3: An example of the LDC GB–English dictionary.

The LDC has made available for research two fairly large bilingual wordlists: GB–English dictionary and English–GB dictionary. In our experiments, we used the GB–English dictionary, which contains 128,366 entries. As shown in Figure 2.3, each Chinese term is followed by English equivalents separated by slashes. The translations may be synonymous with each other, or correspond to the different senses of the Chinese term, as for example line 8 in Figure 2.3. Within a translation, there may be additional clarification or explanation enclosed by parentheses (line 6).

2.3.3 Chinese word segmentation

In contrast to English and other European languages, Chinese text does not have a natural delimiter between words. As a consequence, word segmentation is an essential issue in Chinese–English translation processing. Chinese word segmentation is a process of identifying word boundaries in text and has been widely studied. Accurate segmentation is challenging due to the fact that in many cases a Chinese character can be either a term by itself or apart of a compound terms. The word segmentation problem in Chinese has been extensively researched in the past decade ([Brent and Tao, 2001; Ge et al., 1999; Jin, 1994; Peng and Schuurmans, 2001; Dai et al., 1999; Teahan et al., 2000]). There are two traditional approaches to Chinese segmentation: a lexical rule-based approach and a statistical approach.

In the lexical rule-based approach, a lexicon containing a large number of Chinese words is predefined and then heuristic methods are utilized to match against the lexicon to segment a Chinese sentence. Maximum matching [Wong and Chan, 1996] is a typical heuristic method. It groups the longest initial sequence of characters, which matches a dictionary entry, as a word starting from the beginning (maximum forward matching) or the end of a sentence (maximum

backward matching). When a word is found, it continues at the next character until all the characters in the sentence have been covered. An alternative to maximum matching is minimum matching [Li and Yuan, 1998]. Using the lexical rule-based approach for segmentation, indexing, and requires a smaller inverted index file and therefore obtains higher efficiency. It is also easy to incorporate linguistic knowledge in the retrieval system. The performance of the lexical rule-based approach depends highly on the completeness of the coverage of the adopted lexicon. It is impossible to include all the Chinese words (known words and unknown words) in a dictionary because the set of words is open-ended [Chen et al., 1997]. Building a lexicon is expensive and time-consuming.

In the statistical approach, the lexical statistics of the Chinese characters in corpora are used to mark the boundaries of words. Sproat and Shih [1990] extracted bi-grams with the highest value of mutual information in the sentence recursively until no other bi-grams can be extracted. Their technique was later evaluated by Chen et al. [1997], who found that it performed better than maximum matching. However, their technique can only segment uni-grams and bi-grams. Chien [1997] developed a PAT tree-based method for segmentation. All of the lexical patterns regardless of pattern length are first extracted. A mutual information-based filter algorithm is then used to filter out the character string in a PAT tree. The performance is good but building a PAT tree is time-consuming and large space overhead is required. Yang et al. [2000] developed a boundary detection technique. It detects the segmentation points by using a threshold and abrupt changes in the values of mutual information between adjacent characters in a Chinese sentence. High accuracy and efficiency is achieved. Moreover, it is not restricted to only uni-grams and bi-grams, although both use mutual information for measuring the association between adjacent characters. Other researchers [Yang and Li, 2005; Ren, 2001] have also developed hybrid techniques based on a lexical rule-based and statistical approaches.

The maximum forward matching is adopted in our query translation process. The major advantage of maximum matching is its efficiency while its segmentation accuracy can be expected to achieve around 95% [Wong and Chan, 1996]. We compiled a segmentation dictionary with $133,599$ entries using the two translation dictionaries described in Section 2.3.2, and updated it using our OOV term translation techniques (see Chapter 3) at run-time. However, because there is no standard definition of word boundaries in Chinese, it is difficult to obtain word segmentation accuracy of 100%. High precision segmentations is not the focus of this thesis. Instead we aim to evaluate the effectiveness of our query translation process as long as the errors caused by word segmentation are reasonably low. However, we note that the OOV term translation technique,

which we have developed to identify Chinese OOV terms and extract English translations through web mining, could enhance the coverage of segmentation dictionary and thus improve segmentation accuracy in our query translation process.

2.4 IR Evaluation

The evaluation of IR systems is the process of assessing how well a system meets the information needs of its users. There are two broad classes of evaluation, *system evaluation* and *user-based evaluation* [Voorhees, 2002b]. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system can rank documents. Since the goal is to determine how well an IR system meets the users' information needs, user-based evaluation would seem to be preferable over system evaluation. However, user-based evaluation is expensive and difficult to do correctly. A properly designed user-based evaluation must use a sufficiently large, representative sample of actual users of the IR system; each of the systems to be compared must be equally well developed and completed with an appropriate user interface; each tester must be equally well-trained on all systems. Such considerations lead IR researchers to often use the less expensive system evaluation.

This thesis uses a *system evaluation* to compare the retrieval results of monolingual IR and CLIR systems, and thus investigate the effects of our query translation process on retrieval performance. System evaluation is, by design, an abstraction of the retrieval process that equates good performance with good document rankings. The abstraction allows experiments to control some of the variables that affect retrieval performance thus increasing the power of comparative experiments. These laboratory tests are much less expensive than user-based evaluations while providing more diagnostic information regarding system behavior.

IR research has a well established tradition of comparing the relative effectiveness of different retrieval approaches. A paradigm for system evaluation was first introduced in the Cranfield experiments [Cleverdon, 1991; 1997]. The Cranfield paradigm has been the dominant experimental IR model for four decades, and is the model used in evaluation efforts such as TREC, CLEF, and NTCIR.

2.4.1 IR system evaluation

System evaluation requires a test collection consisting of three distinct components: a collection of documents, a set of queries or topics (the statements of information need), and a set of relevance judgments. The relevance judgments are a list of which documents should be retrieved for each

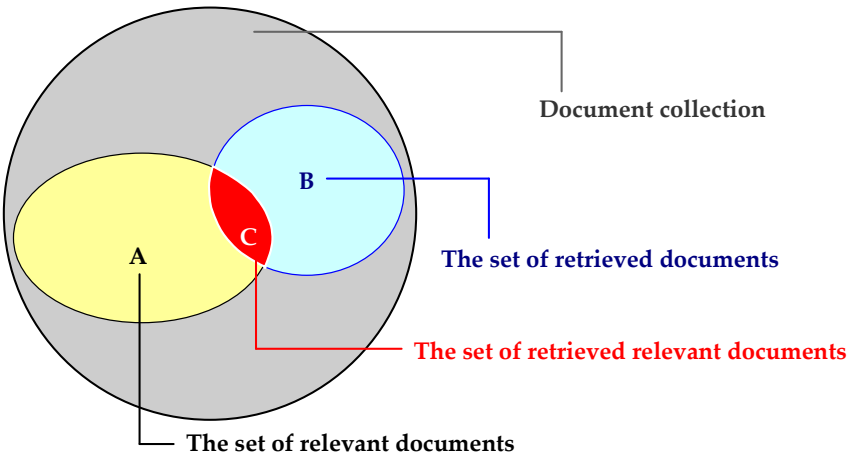


Figure 2.4: Recall and precision definitions.

topic. For an experimental run, the set of queries is used by the retrieval system to retrieve a ranked list of documents from the set of test documents. Based on the retrieved documents and the relevance judgments, evaluation measures concerning retrieval effectiveness can be calculated.

Evaluation measures

Retrieval runs on a test collection can be evaluated in several ways. The most commonly used evaluation measures are *recall* and *precision*. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. For example, a certain document is either relevant to a query or it is not. (The unrealistic binary view of relevance is of course one of the major criticisms of system evolutions.) The same document can either be retrieved or not. As shown in Figure 2.4, recall and precision are defined as:

$$Recall = \frac{number\ of\ relevant\ documents\ retrieved}{total\ number\ of\ relevant\ documents} = \frac{C}{A}$$

$$Precision = \frac{number\ of\ relevant\ documents\ retrieved}{total\ number\ of\ retrieved\ documents} = \frac{C}{B}$$

Specifically, precision measures the ability of a system to retrieve relevant documents and nothing else, while recall measures the ability of a system to retrieve all relevant documents. These two measures, either individually or combined, form the basis for most system evaluation measures.

In the TREC, CLEF, and NTCIR evaluations, there are serveral performance measures typically used in the IR literature:

Recall at 1000 After retrieving 1000 documents, how many of the known relevant documents were found? This is a set-based measure over a group of 1000 retrieved documents, ignoring rank.

Interpolated recall-precision averages (11-point precision) Each recall-precision average is computed by summing the interpolated precisions at the specified recall cutoff value (denoted by $\sum P_\lambda$, where P_λ is the interpolated precision at recall level λ) and then dividing by the numbers of queries — $\sum_{i=1}^{N}(P_\lambda)/N$. (where $\lambda = \{0.0, 0.1, 0.2, ..., 1.0\}$, and N is the total number of queries.)

Average precision Non-interpolated for all relevant documents (averaged over queries). The average of the precision value obtained after each relevant document is retrieved. This measure reflects performance over all relevant documents and rewards systems that ranks relevant documents highly.

Precision at 5, 10, 15, 20, 30, 100, 200, 500, 1000 documents Precision after n documents have been retrieved. Precision at 10 measures how well a system returns only relevant documents from the viewpoint of a web user who typically only looks at the first 10 document retrieved.

R-precision Precision after R documents have been retrieved, where R is the total number of relevant documents for the query. This measure de-emphasizes exact ranking and favors high-precision systems.

In this thesis, we use the average precision, the recall at 1000 (Recall), and the precision at 10 (P@10) to measure our retrieval experiments.

Significance tests

An IR experiment is affected by sampling error. Therefore, to compare the performance of different techniques, we must decide whether the observed difference in performance is statistically significant. For this reason, we use statistical hypothesis testing.

All hypothesis-testing procedures can be broadly described as either non-parametric or parametric. In a parametric test an underlying distribution is assumed (for example, the data analyzed

has a normal distribution), and sample statistics are obtained to estimate the population parameter. Non-parametric methods provide an alternative series of statistical methods that require very limited assumptions to be made about the data. Non-parametric methods have several advantages, or benefits:

- Permit the solution of problems that do not involve the testing of population parameters.
- They may be used on all types of data — qualitative data (nominal scale), data in rank form (ordinal scale), and truly quantitative data (interval and ratio scale).
- They are generally easy to apply and quick to compute when the sample size is small. Sometimes they are as simple as just counting how often some feature appears in the data.
- Depending on the particular procedure they may be almost as powerful as the corresponding parametric procedure when the assumptions of the latter are met, and when this is not the case, they are generally more powerful.

The Wilcoxon signed-rank test, also known as the Wilcoxon matched pairs test, is a non-parametric test used to test the median difference in paired data and determine whether or not difference is statistically significant. In order to compare two systems we look at the set of differences in the retrieval effectiveness values for each of the topics. We assume that the paired differences are independent, each paired difference comes from a continuous distribution that is symmetric, and the paired differences all have the same median. The Wilcoxon signed-rank test consists of six basic steps:

1. Rank the differences without regard to the sign of the difference. All pairs with equal absolute differences get the same rank.
2. Ignore all zero differences.
3. Affix the original signs to the rank numbers.
4. Sum all positive ranks and all negative ranks to give S and determine the total number of differences (denoted by n).
5. Calculate σ^2, where $\sigma^2 = n(n+1)(2n+1)/6$.
6. Systems are different if $S > z(\alpha)\sigma^2$, where α is the significance level pre-specified, $\alpha = 0.05$ means we admit that we have a 5% chance of making a mistake. $z(\alpha)$ is the critical z-value determined by the significance level α.

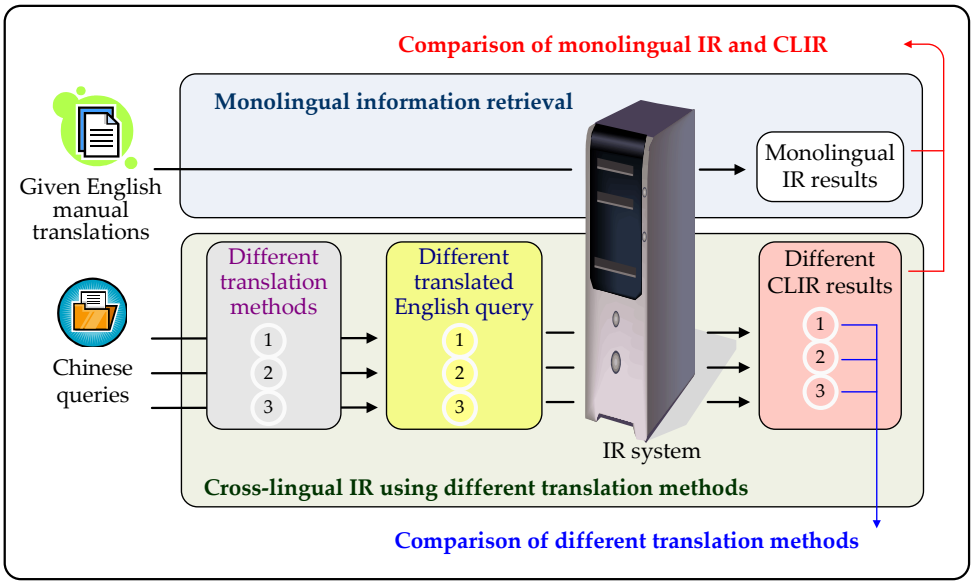


Figure 2.5: Chinese–English CLIR system evaluation methodology.

In general, the smaller the difference between any two systems the larger the sample size that is needed to identify a difference.

2.4.2 CLIR evaluation methodology

In monolingual IR experiments, researchers commonly test different IR systems or different retrieval methods using same test queries and documents, and thus compare the retrieval results generated by different systems or different retrieval methods. The same practice is followed in CLIR experiments when comparing different CLIR systems.

However, in order to directly compare the performances between a CLIR system and its monolingual counterpart, CLIR experiments use different test queries — the given manual translations of the test queries and the system translated queries. The monolingual IR is usually used as the baseline for judging the degree to which query translation process degrades retrieval performance in CLIR. If the problems of OOV translation and translation disambiguation can be sufficiently reduced, we could expect the effectiveness of our CLIR system to be comparable to that of monolingual IR system. To measure the performance of different translation methods, we use the different translations generated by each of the translation methods (as shown in Figure 2.5).

<top>

<num> Number: CH2
<E-title> Communist China's position on reunification
<C-title> 中共对于中国统一的立场

<E-desc> Description:
China, one-nation-two-systems, Taiwan, peaceful reunification, economic and trade cooperation,
cross-strait relationship, science and technology exchanges

<E-narr> Narrative:
A relevant document should describe how China wishes to reach reunification through the implementation of "one-nation-two-systems." If a document merely states a foreign nation's support of China's sovereignty over Taiwan or discusses trade cooperation as well as cultural
and technical exchanges between China and a country other than Taiwan, then the document is irrelevant.

<C-desc> Description:
中国，一国两制，台湾，和平统一，经贸合作，两岸关系，科技、文化交流

<C-narr> Narrative:
相关文件必须提到中共如何经由实现一国两制来达到台湾与大陆统一的目的.如果文件只是外国政府重申支持中共对台湾拥有主权或提到中共与其他国家之经贸、科技、文化交流，则为不相关文件.

</top>

Figure 2.6: TREC Topic numbered CH2.

2.4.3 Test collections

Evaluation experiments in this thesis employed standard test collections, including TREC 5 and 6 Chinese–English track and NTCIR 4 and 5 English–Chinese evaluation collections.

The TREC 5 and 6 Chinese test collection contains 164,789 Chinese documents (170 MB, in GB encoding) of articles drawn from the *People's Daily newspaper* and the *Xinhua newswire*, the associated 54 English topic, and a set of relevance judgments. Each topic consists of three sections: *title*, *description*, and *narrative*. A sample TREC topic is shown in Figure 2.6 and a sample TREC documents is shown in Figure 2.7. In Section 3.3.4, we run a set of experiments to measure the ability of our OOV term translation technique to find appropriate Chinese translations of English OOV terms on these two collections.

<DOC>
<DOCID> CB012018.BFJ (357) </DOCID>
<DOCNO> CB012018-BFJ-357-6 </DOCNO>
<DATE> 1994-04-18 13:02:02 (33) </DATE>
<TEXT>
<headline> 南非因卡塔自由党推迟举行示威游行 </headline>

<s>
新华社约翰内斯堡４月１７日电南非因卡塔自由党青年组织执委会成员马托贝拉今天在记者招待会上宣布，该组织决定推迟举行原定１８日在这里举行的示威游行。 </s>

<s> 但马托贝拉说，群众行动仍将举行，日期将于明日宣布。 </s>
<s> 他还说，示威活动是和平的和非暴力的。 </s>

<s>
因卡塔自由党发言人１５日说，该党支持者将于１８日至２２日在约翰内斯堡举行游行示威，要求推迟大选和取消对纳塔尔地区的紧急状态，并悼念该党在上月政治骚乱中的死亡者。 </s>

<s>
南非警方不允许因卡塔自由党计划中的示威游行，因为约翰内斯堡已被宣布为动乱地区。 </s>

<s> 南非总统德克勒克日前也要求因卡塔自由党放弃示威游行的计划。 </s>

<s>
３月２８日，数万名因卡塔自由党的支持者曾在约翰内斯堡游行示威，从而引发了大规模政治骚乱，造成５０余人死亡，４００多人受伤。 </s>
<s> （完） </s>

</TEXT>
</DOC>

Figure 2.7: TREC Document numbered CB012018-BFJ-357-6. (In GB encoding scheme)

The English documents used in the NTCIR−4 CLIR task contain 347, 376 news articles between 1998 to 1999 collected from different news agencies of different countries. There are 58 Chinese topics for evaluation. The NTCIR−5 English collection contains 259, 050 news articles from 2000 to 2001 and 49 Chinese topics. The format of the NTCIR documents is basically the same as in

```
<DOC>

<DOCNO>XIE19980101.0003</DOCNO>
<LANG>EN</LANG>
<HEADLINE> Turkey Urged to Change Position Towards Greece </HEADLINE>
<DATE>1998-01-01</DATE>

<TEXT>
<P>
ATHENS, December 31 (Xinhua) -- Greek Defense Minister Akis Tsohatzopoulos
today called on neighboring Turkey to change its stance towards Greece.
</P>
<P>
In his new year message, Tsohatzopoulos stressed that aggressiveness did not
help Turkey's relations with the European Union (EU), and Turkey should give
serious consideration to the problems noted by the 15-nation bloc.
</P>
<P>
He pointed out that if Turkey realised this in the future, it would greatly help to
improve the current climate.
</P>
<P>
Earlier this month, the EU decided not to include Turkey on the list of nations
waiting to join the bloc, on the grounds of Turkey's "poor human rights and bad
relations with Greece."
</P>
<P>
After that, Turkey announced the suspension of its political dialogue with the EU
and increased its pressure upon Greece.
</P>
<P>
Tension between the two neighboring countries intensified recently with the
mutual expulsion of diplomats and violations of Greek airspace by Turkish jet
fighters.
</P>
</TEXT>

</DOC>
```

Figure 2.8: NTCIR English document numbered XIE19980101.0003.

the TREC or CLEF collections, plain text with SGML-like tags. A sample English document is shown in Figure 2.8. An NTCIR Chinese topic contains four parts: *title*, *description*, *narrative*, and *key words* relevant to the whole topic. A sample Chinese topic is shown in Figure 2.9. The relevance judgements provided by NTCIR are at two levels — strictly relevant documents known as *rigid relevance*, and documents that are likely to be relevant, known as *relaxed relevance*. We use only *rigid relevance* in our results.

<TOPIC>
<NUM>022</NUM>
<SLANG>KR</SLANG>
<TLANG>CH</TLANG>
<TITLE>合法經營，起亞汽車，意見</TITLE>
<DESC>檢索包含專家對起亞汽車合法經營意見的文章。</DESC>
<NARR>
<BACK>
因爲起亞汽車公司的破產被指爲導致韓國經濟危機的原因，在如何處理此公司的破產問題上引起了全國極大的爭議。</BACK>
<REL>
有關政府官員與專家對處理韓國起亞汽車合法經營程序的意見與批評，和涉及此事件相關問題之文章爲相關。僅包含法律與管理事實之文件爲不相關。
</REL>
</NARR>
<CONC>起亞汽車，合法經營</CONC>
</TOPIC>

(a) NTCIR–4 Chinese topic numbered 022

<TOPIC>
<NUM>022</NUM>
<SLANG>KR</SLANG>
<TLANG>EN</TLANG>
<TITLE>Legal Management, Kia Motors, Opinion</TITLE>
<DESC>Search for articles with professional opinions about Kia Motor's legal management</DESC>
<NARR>
<BACK>
Because the insolvency of Korea's Kia Motors Corporation is indicated as one of the main causes of Korea's economic crisis, there was serious national controversy in how to handle the insolvency of the company.</BACK>
<REL>
Documents that describe opinions or the criticism made by government bureaucrats or professionals about the process to legally manage Korea's Kia Motors and the problems involved are relevant. Documents that simply include legal and administrative facts are irrelevant.
</REL>
</NARR>
<CONC>Kia Motors, legal management</CONC>
</TOPIC>

(b) The given English translation of the NTCIR–4 Chinese topic 022

Figure 2.9: NTCIR topic numbered 022. (The Chinese topic is in BIG5 encoding scheme.)

In our Chinese–English CLIR experiments (see Chapter 5), we translate the titles and descriptions of the NTCIR Chinese topics into English using the combinations of the translation disambiguation and OOV term translation techniques we developed in Chapter 3 and Chapter 4, and then use the translated English queries to retrieve the documents from the NTCIR English document collection. We note that the average length of the titles is around 3.5 terms, which approximates the average length of web queries.

2.5 Summary

In this chapter, we presented an overview of translation approaches to CLIR, such as document translation, query translation, and transitive translation (where there is no direct translation resource between the query language and the document language). In previous studies, transitive translation has been much less effective than query or document translation. Query translation has emerged as the most popular approach to CLIR, because document translation is time-consuming and requires re-indexing of the entire collection each time the automatic translation system is modified to produce a new target collection.

Within the query translation framework, we described three main approaches in CLIR: the use of machine translation, a parallel corpus, or a bilingual dictionary. By far the most commonly used query translation approach is to replace each query term with appropriate translations that are automatically extracted from a bilingual dictionary. Many researchers adopt dictionary-based query translation not only because it is simple and practical, given the wide availability of bilingual dictionaries; but also it has proved to be the most robust for the short queries that are typically entered by web users.

However, there are well known problems with dictionary-based query translation, namely, the phrase translation problem, the term ambiguity problem, and the OOV problem. These problems may result in very poor retrieval performance of the translated queries. We have reviewed previous studies that explored these problems.

There are some inherent linguistic difficulties in Chinese text retrieval, in particular word segmentation and identification of unknown words. Since written Chinese texts lack explicit delimiters between words to indicate boundaries, word segmentation is a prerequisite and the major barrier for Chinese text retrieval. In Section 2.3, we have discussed the characteristics of Chinese text, issues associated with Chinese text retrieval, and practical solutions to the challenges of Chinese text processing.

Finally, we discuss the effectiveness of IR systems. Such an evaluation is usually based on

a test collection and a set of evaluation measures. The test collection consists of a collection of documents, a set of queries, and a set of relevance judgments. For an experimental run, the set of queries is used by the IR system to retrieve a ranked list of documents from the set of test documents. Based on the retrieved documents and the relevance judgments, evaluation measures concerning retrieval effectiveness can be calculated. IR experiments are stochastic experiments affected by random errors. Therefore, to compare the performance of different approaches, we must decide whether the observed difference in performance is statistically significant. It is not sufficient to compare different retrieval approaches by their mean performances, because these can be heavily affected by outliers. Instead, we compare the distributions of the observations. The statistical significance of the performance difference is best checked using non-parametric tests, because these tests make least assumptions on the experimental data. We have described the Wilcoxon signed-rank test that has been widely used in IR in Section 2.4.1. Additionally, we describe our CLIR experimental methodology, including the test collections employed in our evaluation experiments.

In the next two chapters, we will examine the problems of OOV translation and translation disambiguation in the context of Chinese–English CLIR.

Chapter 3

OOV Term Translation

Queries that seek topical information typically contain out-of-vocabulary (OOV) terms that are not found in a translation dictionary, leading to inappropriate translations and consequent poor retrieval performance. For example, the query may concern current affairs, and thus contain new words or translated words that are outside the scope of the translation dictionary; or the query may contain proper nouns — such as brand names, place names or personal names — that are not included in the translation dictionary. Although only some queries contain OOV terms, incorrect translation of such terms almost inevitably leads to disastrous results.

Successful translation of OOV terms is one of the challenges of CLIR. Particular difficulties arise in languages where there are no clearly defined boundaries between words, as is the case with Chinese text. When translating from Chinese to English, a standard first step is to segment the text into words based on a segmentation dictionary. However an OOV term will not be recognized, and will thus be segmented into either smaller sequences of characters or into individual characters. In this case the constituent components may be translated into terms in the target language that have little relationship to the original meaning. When translating from English to Chinese, existing systems are able to detect an English OOV term since no segmentation is required. They are either present in the translation dictionary or not. However when trying to discover appropriate Chinese translations for these terms, segmentation again comes into play, and previous work [Chen et al., 2000] has suffered from inability to correctly identify new terms automatically.

We consider related background work in OOV term translation and explain the ideas behind the extensions we have developed in Section 3.1. In Section 3.2, we describe our algorithm for extracting English translations of Chinese OOV terms; in Section 3.3, we give our algorithm for

Figure 3.1: An example of a Chinese web page containing both English and Chinese text.

extracting Chinese translations of English OOV terms.

3.1 Segmentation Free Translation Extraction

Our approach stems from the observation that new terms, foreign terms, or proper nouns in Chinese web text are sometimes accompanied by the English translation in the vicinity of the Chinese text, as shown in Figure 3.1. By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we are then able to infer the appropriate translation with reasonable confidence. The idea of using the web to search for translations is not new [Chen et al., 2000; Meng et al., 2004], however our technique is segmentation-free and

consequently can extract translations that were previously undetected, or only detected by manual intervention to provide correct segmentation. It is common to find a small amount of English text in Chinese web documents, but extremely rare to find Chinese text in English web documents. We therefore rely on Chinese web documents to extract translations in both directions.

Segmentation causes difficulty in both directions, but different approaches are needed in each case. When looking for English translations of Chinese OOV terms, the first step involves appropriately detecting the Chinese OOV terms. Typically a segmenter is used to determine Chinese word boundaries, and this information would be used to assist in the identification of the Chinese OOV term. The problem is that the Chinese OOV term we are looking for is currently unknown, and thus we have no information about how it should be segmented. For example: suppose the query is "$x_1x_2x_3x_4x_5x_6x_7$" and the correct segmentation is "$x_1x_2x_3 \mid x_4 \mid x_5x_6x_7$". However "$x_5x_6x_7$" is not in any available dictionary, and thus is an OOV term. If we first applied a segmenter, we might obtain "$x_1x_2x_3 \mid x_4 \mid x_5 \mid x_6x_7$". We could be tempted to conclude that two or more characters "x_4x_5" are an OOV term, but that would be wrong. It is unwise to assume anything about the segmentation of a query when trying to identify OOV terms. In previous work [Chen et al., 2000], this problem was overcome by manual intervention to provide appropriate segmentation. Looking for a Chinese translation of an English OOV term is also not straightforward, since typically a number of candidate Chinese character strings are found and must then be segmented. Previous automatic procedures [Meng et al., 2004; Chen and Gey, 2003] have not been particularly successful. Our technique overcomes the difficulties experienced previously by researchers in this area. The details of each procedure are explained in the following sections.

3.2 Chinese OOV Term Detection and Translation

When looking for English translations of Chinese OOV terms, they need to be appropriately detected in the query. We have observed that some systems sometimes do not translate an OOV term at all (such as BabelFish), while others translate character by character using Pinyin, with disastrous results. Many existing systems use a segmenter to determine Chinese word boundaries. However, if the Chinese OOV term is currently unknown, there is no information to indicate how it should be segmented. In other work [Chen et al., 2000], this problem appears to have been overcome by manual intervention to provide appropriate segmentation. However, it is clearly desirable that the segmentation be either automatic or, as in the case of the technique we describe, unnecessary.

When a large corpus of Chinese text is available, it is possible to apply statistical techniques

to identify named entities that are not present in translation dictionaries. Sun et al. [2003] used a trigram stochastic model to detect named entities. Their technique had a success rate of approximately 80%; however, they did not attempt translation. Such a technique is not practical in our situation, as we do not have a Chinese corpus to work with, and in any case it is unlikely that a corpus would contain many of the OOV terms that occur in news and current affairs. The basis of our approach is the observation that most translated English terms tend to accompanied by the original English terms on the web, typically immediately after the Chinese text, but general terms do not. For example, the text might contain "世紀之毒戴奧辛（Dioxin）" where "世紀之毒戴奧辛" is a sequence of Chinese characters and "Dioxin" is the original English term for "戴奧辛". By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we hypothesize that we are then able to infer an appropriate translation with reasonable confidence. In formulating our approach, we also considered English text that was not immediately adjacent to the Chinese query terms. However, we found that such text was only rarely a reliable translation. In some cases we found only a small number of Chinese–English co-occurrences, and our approach proved to be robust in such situations. In summary, our procedure consists of three steps: extraction of the web text, collection of co-occurrence statistics, and translation selection.

3.2.1 Extraction of web text

First, we query the web to identify strings that contain the Chinese query terms and some English text.

1. Use a search engine to fetch the top 100 Chinese documents, using the entire Chinese query. (A side effect of using a reliable search engine is that only good-quality web text is returned. This reduces the likelihood of noisy translations being collected.)

2. For each returned document, the title and the query-biased summary are extracted.

For example, consider a query "北野武導演的電影" (Director Takeshi Kitano's films) composed of four Chinese terms: "北野武" (Takeshi Kitano), "導演" (Director), "的"('s), and "電影" (films), suppose that "北野武" is an OOV term, "的" is a structural particle, and "導演" and "電影" are in-vocabulary terms. We used this query to retrieve a series of titles and query-biased summaries of web text that contain English text, as shown in Figure 3.2, and as can be seen, "Takeshi Kitano" is the commonest English text and "北野武" is the Chinese text most commonly observed in the context.

... 北野武性愛狂想曲【Geetting Any】...
.. 雖然之前還為石井隆的《Gonin》作過序 ..
... 因為我這個只對大衛．林區(David Lynch)等幾位導演或犯罪..
... 歡迎到eWorld 好站...
... ... 導演北野武(Takeshi Kitano) ... ...
... ... 北野武(Takeshi Kitano) ...
... ...導演 北野武 Takeshi Kitano ... ...
... 他是BeatTakeshi ...
... 菊次郎的夏天Kikujiro ..
... 【盲劍俠】ZATOICHI-劍客側寫...
... 【盲劍俠】ZATOICHI-劍客側寫...
...【那個凶暴的男子】(Violent Cop) ...
... 【3比4X十月】(Boiling Point) ...
... 《３比４Ｘ十月》（BOILING POINT 1990）...
...《奏鳴曲》（SONATINE 1993）... 《恣在年少》（KIDS RETURN 1996）...
... 北野武(1948~\--) ...
... 導演：北野武Takeshi Kitano ... 演員：北野武Takeshi Kitano ...
... 日本性格小生淺野忠信(Tadanobu ASANO) ...
... The Spinning Image「視覺暫留」作者：Daniel Auty ...
... ... 北野武（ Takeshi Kitano ） ... ...
... ... 導演(Director) ...
... 電影DVD ... ...
... ... 導演Takeshi Kitano ...
... ... 電影: CHARLIE AND THE CHOCOLATE FACTORY ...
... ... 北野武(Takeshi Kitano) ... ...

Figure 3.2: Web text retrieved using "北野武導演的電影" (Director Takeshi Kitano's films).

3.2.2 Collection of co-occurrence statistics

We then collect co-occurrence information from the data we obtained. Although much English text is present in web pages written in Chinese, not all of it is useful in OOV translation. We only consider the English text that occurs immediately after the Chinese query substrings, because, if such an English term occurs at a high frequency, it almost invariably serves as the translation of that Chinese string.

1. Where English text occurs, check the immediately preceding Chinese text to see if it is a substring of the Chinese query.

2. Collect the frequency of co-occurrence of each distinct English string and all Chinese query substrings that appear immediately prior.

e	f_e	c_j	$\|c_j\|$	$f(e, c_j)$
Takeshi Kitano	8	北野武	3	7
		導演	2	1
Director	1	導演	2	1
1948	1	北野武	3	1

Table 3.1: The co-occurrence frequency of English terms and Chinese query substrings

For each distinct English string e with the frequency f_e, we obtained a group of associated Chinese query substrings c_j with the length $|c_j|$ and the co-occurrence frequency $f(e, c_j)$. Extending the example from Figure 3.2, this information is summarized in Table 3.1.

3.2.3 English translation selection

Incorrect translations can, as discussed earlier, greatly degrade effectiveness. For this reason, we select only the best translation for a Chinese OOV term. The selection algorithm is shown as follows:

1. Select the English text e with the highest f_e, since the remaining English text that occurs with the highest frequency is more likely to provide the correct translation compared to other text with the lower frequency.

2. For this English text e, select the associated Chinese query substring c_j with the highest $f(e, c_j)$. In the event of a tie we use $|c_j|$ to discriminate.

3. If the selected Chinese query substring c_j cannot be found in the Chinese segmentation dictionary, we treat it as OOV term and add it into the Chinese segmentation dictionary and (c_j, e) into the translation dictionary.

In this case, "北野武" is identified as a Chinese OOV term and "Takeshi Kitano" is extracted as its English translation from Table 3.1.

A given Chinese term may have more than one English transliteration. For example, "Osama" and "Usama" are transliterated into the same Chinese term "奧薩瑪". However, this phenomenon is rare. Our methods tend to choose the most common form.

3.3 English OOV Term Detection and Translation

Our work builds on the work of Chen and Gey [2003]. In their translation extraction process, each English OOV term is submitted as a query to *Yahoo!Chinese* in BIG5 encoding. The top 200 result entries are then segmented into words using a dictionary-based longest matching method. For each line containing the English query word or phrase, they consider the five Chinese "words" immediately before and after the English word or phrase, and use a weighting scheme to select the top m of these as the best translation, where m is the number of English terms. Since the Chinese word being searched for is currently unknown, there is no information as to how it should be segmented, and thus segmentation errors may occur, leading to an incorrect translation extraction. When this occurs the retrieval effectiveness is inevitably substantially degraded. In an example provided, 7 out of 17 extracted translations exhibited this problem. Another problem is that sometimes the correct Chinese translation may occur some distance from the English OOV term, for example "TAFE 的全称是技术与继续教育学院", where "技术与继续教育学院" is the Chinese translation of the English OOV term "TAFE".

In contrast to Chen and Gey [2003], our technique does not rely on a prior segmentation and is based on the consideration of every possible Chinese substring occurring adjacent to the English OOV term. In our experiments, we have found this procedure to be free from segmentation error in translation extraction. We describe below our process to automatically extract the Chinese translations of English OOV terms from the web.

3.3.1 Extraction of web text

First, we extract strings that contain the English OOV term and some Chinese text from the web using the English OOV term as the query. The basic method of web text extraction is the same as in Section 3.2.1.

3.3.2 Collection of co-occurrence statistics

We then collect statistical information from the data we obtained in three steps.

1. Where the English OOV term occurs, we accumulate its frequency f_{oov} and collect W Chinese characters immediately before as C_{left} and immediately after as C_{right}.

2. Since we want to use a process that does not rely on segmentation, we start by considering all substrings in C_{left} and C_{right} as candidate Chinese translations, and collecting the frequency f_n and the length $|c_n|$ of each Chinese substring c_n.

3. The candidate Chinese translations are ranked based on the likelihood of being the correct translation. Generally we prefer candidate translations that occur more frequently over those that occur less frequently, and prefer long candidate translations over short ones. However the natural distribution is that shorter candidate translations occur more frequently. The score of each c_n is calculated using:

$$S_{c_n} = \alpha \times \frac{|c_n|}{W} + (1 - \alpha) \times \frac{f_n}{f_{oov}}$$

From our experiments with post hoc tuning, we determined that $W = 20$ and $\alpha = 0.25$ provides the best combination of frequency and length, a value that proved to be robust across our experiments.

For example, we use the English OOV term "TAFE" as the query to retrieve a series of titles and query-biased summaries of web text. The frequency and the length of the top-ranked Chinese substrings are summarized in Table 3.2, which we refer to in the remaining steps.

3.3.3 Chinese translation selection

From these top-ranked Chinese substrings in Table 3.2, we select the most appropriate translations in the following manner:

1. We exclude any candidate translation that is already in the translation dictionary from the candidate translation set. In this case, "学院" (college), "教育" (education), "澳大利亚" (Australia), and "技术" (technology) are removed from the translation candidate set.

2. We discard any candidate translation from the candidate translation set if (a) it is a substring of any of those translations excluded in the previous step and (b) it has a lower S_{c_n}. The candidate translations "大利亚", "澳大利", "利亚", "澳大", and "大利" are filtered out in this step.

3. Any candidate translation that does not occur in the document collection is excluded from the candidate translation set.

4. We then select the candidate translation with the highest S_{c_n} from the candidate translation set. In this case, "技术与继续教育学院" is extracted as the best Chinese translation of the English OOV term "TAFE".

n	c_n	$\|c_n\|$	$f(c_n)$	S_{c_n}
1	学院	4	110	0.477461
2	教育	4	98	0.430829
3	澳大利亚	8	75	0.391451
4	大利亚	6	75	0.366451
5	澳大利	6	75	0.366451
6	利亚	4	76	0.345337
7	澳大	4	76	0.345337
8	大利	4	75	0.341451
9	技术与继续教育学院	18	25	0.322150
10	技术	4	65	0.302591
11	技术与继续教育学	16	25	0.297150
12	术与继续教育学院	16	25	0.297150
13	技术与继续教育	14	28	0.283808
14	术与继续教育学	14	25	0.272150
15	与继续教育学院	14	25	0.272150
16	技术与继续教	12	28	0.258808
17	术与继续教育	12	28	0.258808
18	继续教育学院	12	26	0.251036
19	与继续教育学	12	25	0.247150
20	术与继续教	12	28	0.233808

Table 3.2: The statistical information for the 20 *top-ranked candidate Chinese translations collected from the extracted web text for the English OOV term "TAFE".*

3.3.4 English OOV term translation experiments

In this section, we describe the experimental setup for retrieving Chinese documents using English queries. The goal of this set of experiments is to measure the ability of our English OOV term translation technique to find appropriate Chinese translations of English OOV terms.

The test collection used in this task is the TREC 5 and 6 Chinese collections (as described in Section 2.4.3). We used the maximum forward matching (see Section 2.3.3) to segment the document collection. Not every query contains English OOV terms and such queries are not affected by the OOV problem. As we are particularly interested in OOV, we have only selected the queries containing English OOV terms. In order to mimic typical web queries, we decided to use only the *title* of topics as queries. To maximize the number of test queries, we augmented some

Topic Number	Detected English OOV terms	Extracted Chinese Translations	Given Chinese Translations
2	reunification	和平统一	统一
2	cross-strait	海峡两岸关系	两岸
3	Daya Bay	大亚湾	大亚湾
3	Qinshan	秦山	秦山
7	Dongsha Islands	东沙群岛	东沙群岛
7	Xisha Islands	西沙群岛	西沙群岛
7	Spratly Islands	南沙群岛	南沙群岛
8	Richter	里氏	芮氏
11	Peace-keeping	维和部队	维和
14	HIV	艾滋病毒	(No translation)
21	Peng Dingkang	彭定康	彭定康
21	Reunification	和平统一	统一
28	PSDN	分组交换	分组交换网
31	Castro	卡斯特罗	卡斯特罗
42	Liaohe River	辽河	辽河
42	Haihe River	海河	海河
42	Huaihe River	淮河	淮河
42	Songhua River	松花江	松花江
42	Pearl River	珠江	珠江
46	Sino-Vietnamese	(Not Found)	中越
47	Pinatubo	皮纳图博火山	皮纳图博
47	Subic Bay	苏比克湾	苏比克湾
48	Kuwaiti	科威特第纳尔	科威特
49	Non-Proliferation Treaty	不扩散核武器条约	不扩散核武器条约
49	START	战略武器条约	消减战略武器条约

Table 3.3: Extracted Chinese translations of English OOV terms from TREC 5 and 6 English topics.

otherwise OOV-free topics with English OOV terms from the *description* and *narrative* sections (see Table 3.3). This provided a total of 14 queries.

We compiled a GB–English translation dictionary for this set of experiments using 3 dictionaries for the Chinese–English query translation: an English–GB wordlist and a GB-to-English wordlist

from the LDC (as previously described in Section 2.3.2), and the CEDICT[1] GB–English dictionary. Kwok [2000] showed that the Chinese-to-English wordlist can be considered as both a phrase and word dictionary for English–Chinese cross language retrieval, and is preferable to the English-to-Chinese version in terms of phrase translations and word translation selection. Our translation dictionary contains 128,527 entries including 19,081 multi word phrases that were used for English phrase identification and translation.

Experimental design

In our first run we used the given Chinese queries without any of the Chinese equivalents of the English OOV terms (C–C), and used this to compare the performance of the translated English queries without any English OOV terms (E–C). This allowed us to test the basic effectiveness of our system using the translation disambiguation technique, without regard to the OOV problem. We then manually added the Chinese equivalents of the English OOV terms to the Chinese queries (CO–C), and used this as a further baseline to test the ability of our technique to automatically find the appropriate Chinese translations of the English OOV terms. This was done by adding the English OOV terms to the English queries and using our system to translate and then retrieve Chinese documents (EO–C). Our English-Chinese CLIR experiments used the MG [Witten et al., 1999] search engine.

Results and discussion

Table 3.4 shows a comparison of the interpolated recall-precision averages for the English–Chinese CLIR experimental results. Without any English OOV terms, our translated queries achieved 86.7% of the monolingual result. The underlying performance of our query translation system was affected by the following two factors. First, some of the English query translations provided by the TREC organizers did not precisely parallel the original Chinese queries; for example, in CH47, "菲律宾 (Philippines)" is missing from the English query; in CH21, there is no exact English equivalent of "回归中国(return to China)" given in the English query. Second, some translations provided by the translation dictionary are inappropriate in the given context; for example, in CH28, the English term "cellular phone" is translated into "汽车电话" (car phone), where the given Chinese equivalent is "移动电话". Additionally, in CH47, "impact" is translated into "冲击" (strike), where the given Chinese equivalent is "后果". When the translated English OOV terms were added, we achieved 77.1% of the monolingual result.

[1] http://www.mandarintools.com/download/cedictgb.zip

Recall	**C-C**	**E-C**	**CO-C**	**EO-C**
0.00	0.4850	0.4639	0.6261	0.5698
0.10	0.2839	0.2361	0.4992	0.3928
0.20	0.2208	0.1953	0.4160	0.3561
0.30	0.1828	0.1694	0.3431	0.3030
0.40	0.1411	0.1369	0.2985	0.2561
0.50	0.1104	0.1055	0.2625	0.2121
0.60	0.0824	0.0718	0.2255	0.1797
0.70	0.0624	0.0409	0.1933	0.1291
0.80	0.0339	0.0276	0.1293	0.0614
0.90	0.0089	0.0077	0.0828	0.0280
1.00	0.0000	0.0000	0.0082	0.0057
Average precision	0.1284	0.1113	0.2610	0.2013
% Monolingual	–	86.68	–	77.13

Table 3.4: Effect of English OOV term translation on English–Chinese CLIR effectiveness, recall-precision table showing 11-point interpolated recall-precision averages and average precision for TREC 5 and 6 queries. (Data consists of the TREC 5 and 6 Chinese collections and 14 English topics with English OOV terms. Topic 2, 3, 7, 8, 11, 14, 21, 28, 31, 42, and 46–49.)

Besides the two factors discussed above, we were not able to automatically find the most appropriate Chinese translations for every English OOV term. We failed to find the translation of "Sino–Vietnamese" in CH46, and thus did not obtain any improvement in retrieval effectiveness. In CH48, the correct translation of "Kuwaiti" was lost because we did not consider any translation that is already in the translation dictionary. This is a shortcoming of our extraction algorithm, which assumes that no English OOV term shares the same Chinese translation with any English term in the translation dictionary. In CH49, we extracted an inappropriate translation of "START treaty", because there is currently no single dominant accepted Chinese translation for this term; "削减战略武器条约", "削减进攻性战略武器条约", and "限制战略武器条约" are used interchangeably. It can be seen in Table 3.4 that successful translation of English OOV terms results in a 80% improvement in average recall-precision. This is partly because we have specifically selected the queries known to contain English OOV terms. While the improvement in a heterogenous set of queries would be more modest, the results show that first, not being able to translate OOV terms leads to significant loss in retrieval performance for such queries; and second, our technique is

effective in automatically finding appropriate translations of English OOV terms. Although we were only able to automatically find appropriate translations for 88% (22 out of 25) of English OOV terms, this is a considerable improvement on previous work in this area [Chen et al., 2000], where 72% (8 out 11) translations required manual segmentation correction in order to be processed correctly.

As was the case with our Chinese–English experiments, the final sample size was somewhat small. To further test the robustness of our technique, we collected 50 English OOV terms from news web sites and applied our Chinese translation extraction technique. However, since our technique requires a corpus to provide disambiguation, we were not able to carry out the final step of our translation extraction procedure, namely using the corpus to select the most appropriate translation from a set of candidate translations. Table 3.5 and Table 3.6 show set of candidate translations for each English OOV term. It can be seen that 14 of the English OOV terms have a single correct Chinese translation. A further 31 English OOV terms have at least one correct translation in the candidate set, while our technique failed to find correct translations in 5 instances. In our experiments using the TREC 5 and 6 Chinese data, we arrived at similar situation at the penultimate stage, but with the aid of corpus-based disambiguation were able to select the most appropriate translation in the final phase. We were concerned that the English OOV terms extracted from Chinese web pages might only pertain specifically to Chinese news. However, we note that we were able to find Chinese translations for 96% of English OOV terms from TREC 5 and 6. This gives us some confidence that the technique should at least have general applicability to news based queries.

3.4 Summary

A major difficulty in CLIR is the detection and translation of OOV terms; for OOV terms in Chinese, another difficulty is segmentation. In this chapter, we have investigated the OOV problem as it applies to Chinese–English and English–Chinese CLIR. Previous work has either relied on manual intervention or has only been partially successful in solving this problem. We developed techniques to dynamically discover translations of OOV terms through the mining of web text. Our techniques do not rely on prior segmentation and are thus free from segmentation error. To detect and translate an OOV term in a Chinese query, we use the entire Chinese query to fetch web documents written in Chinese. We then collect the English text that is preceded by any substring of the original Chinese query. For each distinct English string, we obtain a group of associated Chinese query substrings and the occurrence frequency. Finally, we identify the Chinese

	English OOV terms	Chinese Translation Candidate Set
1	Pervez Musharraf	穆沙拉夫| 总统佩尔韦兹穆沙拉夫
2	Shiite Muslim	在联合国安理会讲话全| 联合国安理会讲话全文
3	Ali al-Sistani	赛义德阿里阿尔| 最高称号赛义德阿里
4	Paul Bremer	拉克| 新浪教育新浪网
5	Avian flu	禽流感| 对抗香港禽流感的疫苗
6	Hambali	汉巴里| 在东南亚地区组织恐怖
7	Mad Cow	疯牛病
8	Nintendo	任天堂
9	Carlsberg	嘉士伯
10	Credit Lyonnais	信贷银行| 里昂信贷银行
11	Osama bin Laden	拉登| 本拉登
12	John Howard	华德| 澳大利亚总理
13	Saddam Hussein	萨达姆
14	Kofi Annan	联合国秘书长| 联合国秘书
15	Hezbollah	真主党| 黎巴嫩武装恐怖组织
16	NASD	证券交易商协会
17	PricewaterhouseCoopers	永道| 普华永道
18	SARS	非典
19	Matrix Reloaded	客帝国| 重装上阵
20	Martha Stewart	尔特| 玛莎斯图尔特
21	Tour de France	环法| 环法自行车赛
22	NMD	美国国家导弹防御系统
23	Mars rover	火星探测器
24	Blaster	冲击波| 病毒专杀工具
25	Forest Gump	欧美电影音乐| 阿甘正传

Table 3.5: Extracted Chinese translations of the English OOV terms randomly selected from news web sites. (Terms 1-25)

OOV term and extract its best English translation using our selection algorithm. To translation an English OOV term, we again fetch Chinese web documents, using the English OOV term as query. We consider every possible Chinese strings occurring adjacent to the English OOV term. By applying simple statistical techniques based on frequency and length analysis, we extract the best Chinese translation.

In experiments with our English OOV term translation technique on several collections and a

	English OOV terms	Chinese Translation Candidate Set
26	DJIA	道琼斯工业
27	NAS	网络存储器
28	Land Rover	陆虎
29	Kim Clijsters	克里| 克里斯特尔斯
30	Likud Party	利库德| 利库德集团
31	Lord of the Rings	魔戒| 魔戒首部曲魔戒现身
32	Starbucks	星巴克| 星巴克咖啡
33	Enron	安然公司| 安然
34	Abdullah Gul	外长阿卜杜拉居尔| 总理兼外长阿卜杜拉居
35	Olympus	相机| 奥林巴斯
36	Cappuccino	卡布奇诺| 萧亚轩
37	Espresso	意大利特浓咖啡
38	Mohammad Khatami	哈塔米| 穆罕默德
39	Finding Nemo	海底总动员
40	Arnold Schwarzenegger	瓦辛格| 阿诺德施瓦辛格
41	Rupert Murdoch	默多克| 新闻集团
42	Lancome	兰蔻
43	TAFE	技术与继续教育学院
44	Logitech	罗技| 鼠标
45	PDP	等离子| 等离子显示器
46	Aopen	建基| 主板|
47	ViewSonic	优派| 优派系列显示器
48	Ariel Sharon	以色列总理沙龙
49	Donald H. Rumsfeld	拉姆斯| 拉姆斯菲尔
50	N-Gage	诺基亚| 游戏手机

Table 3.6: Extracted Chinese translations of the English OOV terms randomly selected from news web sites. (Terms 26-50)

set of terms from news articles, we showed that it is robust and provides a substantial improvement in English OOV term translation quality. This technique can be used to correctly extract Chinese translations of English OOV terms from the web automatically, and thus is a significant improvement on earlier work. The full details of our Chinese OOV term translation experiments will be described in Chapter 5.

Chapter 4

Translation Disambiguation

Amongst the various approaches to CLIR, translation of the query language to that of the document language has been most commonly used, and researchers have applied dictionary-based translation methods with success. However, translation ambiguity is a frequent cause of failure of dictionary-based translation, because many words or phrases in one language can be translated into another language in multiple ways, and sometimes the alternative translations have very different meanings. For example, the Chinese word "素质" can be translated as "diathesis", "quality", or "accomplishment"; the choice of English translation depends on the context in which "素质" occurs.

As described in Section 2.2.3, when using simple dictionary translations without addressing the problem of translation ambiguity, the effectiveness of CLIR can be 60% lower than that of monolingual retrieval [Ballesteros and Croft, 1998]. This problem is particularly severe in view of the observed tendency of web users to enter short queries (often two or three words [Aljlayl and Frieder, 2001; Pu et al., 2002]); it is sometime difficult even for a human to reliably determine the intended meaning from the available context. The dictionary-based translation approaches are prone to error due to the likelihood of selecting the wrong translation of a query term among the alternatives provided by the dictionary.

In this research, we propose an improved disambiguation technique for dictionary-based query translation. Our motivation is that, when two words co-occur in the same query, they are likely to co-occur in the same text windows in the documents; moreover, two words being in close proximity generally provides stronger correlation. We therefore integrate a contextual window and a distance factor into the language model. A second issue we investigate is the corpus used

to provide disambiguation. As described in Section 2.4.1, a test collection consisting of three components — a target collection of documents, a set of queries, and a set of relevance judgments — is used in the TREC, CLEF, and NTCIR CLIR evaluations. It is normal to use the target collection as a language resource for disambiguation in query translation. However, the target collection is likely to contain the correct translations of the set of given queries, since it is known to contain relevant documents for each of the queries. In a realistic web search situation, there is no such a tailored collection available for a given web query. Thus an important issue is how to select an appropriate collection of web documents for query disambiguation, and also how the quality of disambiguation compares to that of using a pre-defined collection corpus.

In Section 4.1, we review existing approaches to translation ambiguity. In Section 4.2, we propose an improved disambiguation technique for Chinese–English query translation. In Section 4.3, we first investigate the effect of context window size and a distance factor on disambiguation performance; second, we explore how well our disambiguation technique performs when the web is used to provide context; and finally, we compare our disambiguation technique to other approaches that use different underlying principles.

4.1 Existing Disambiguation Approaches

Web queries are generally short, and thus sophisticated natural language processing techniques can flounder due to lack of context. However, any dictionary-based technique must address the fact that many words have multiple translations. This disambiguation phase is crucial to the translation result. Various techniques have been proposed to reduce the ambiguity and errors introduced during query translation. Instead of using all possible translations in the machine readable dictionary, researchers have proposed techniques based on term co-occurrence [Ballesteros and Croft, 1998; Gao et al., 2002], term similarity [Adriani, 2000; Maeda et al., 2000], and language modelling [Federico and Bertoldi, 2002]. A different approach to reducing the effect of the ambiguity problem is to combine the results of translating query terms using a general and a domain-specific dictionary, and then use structural tags to indicate the contextual relationship among the resulting terms [Pirkola, 1998].

Ballesteros and Croft [1998] described a technique that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary translation. Their hypothesis is that the correct translation of query terms should co-occur in target language documents where incorrect translations should not tend to co-occur. They demonstrated the effectiveness of translating phrases in Spanish queries into English phrases using terms that co-occur in the English

collections. Compared to the corpus-based methods described in Section 2.2.2, co-occurrence appears to be significantly better at disambiguation. Other studies [Hiemstra and de Jong, 1999; Kwok, 2000; Jang et al., 1999] also used similar approaches to select the best translations. Building on the work of Ballesteros and Croft [1998], Gao et al. [2002] observed that the correlation between two terms is stronger when the distance between them in a corpus is shorter. They successfully extended the co-occurrence model by incorporating a distance factor $D(x, y) = e^{-\alpha(\mathrm{Dis}(x,y)-1)}$. The mutual information (MI) [Church and Hanks, 1990] between term x and y, $\mathrm{MI}(x, y)$, is calculated as follows:

$$\mathrm{MI}(x, y) = \log\left(\frac{f_w(x, y)}{f_x f_y} + 1\right) \times D(x, y)$$

where $f_w(x, y)$ is the frequency with which x and y co-occur within a window size of w in the collection; f_x is the collection frequency of x; and f_y is the collection frequency of y. $D(x, y)$ decreases exponentially when the distance between two terms x and y increases, where α is the decay rate, and $\mathrm{Dis}(x, y)$ is the average distance between x and y in the collection. They experimented on the TREC−9 Chinese collection and showed that the addition of the distance factor leads to substantial improvements over the basic co-occurrence model. They were able to achieve 84% of the monolingual effectiveness when using $\alpha = 0.8$ in this decaying co-occurrence model.

Adriani [2000] proposed a translation disambiguation technique based on the concept of statistical term similarity, for selecting the best Indonesian translation of an English term from all possible translations given by a bilingual dictionary. These techniques use a term-similarity matrix built using the statistical term-distribution parameters obtained from the Indonesian corpus for Indonesian terms, and using a subset of their English collections for English terms. The degree of similarity or association-relation between terms is obtained using a term association measure, the *Dice similarity coefficient* [Rijsbergen, 1979], which is commonly used in document and term clustering. The term similarity value between term x and y, $\mathrm{SIM}(x, y)$, is calculated as follows:

$$\mathrm{SIM}(x, y) = 2\sum_{i=1}^{n}(w'_{xi} \times w'_{yi}) \Big/ \left(\sum_{i=1}^{n} w_{xi}^2 + \sum_{i=1}^{n} w_{yi}^2\right)$$

where

w_{xi} = the weight of term x in document i
w_{yi} = the weight of term y in document i
w'_{xi} = w_{xi} if document i also contains term y, or 0 otherwise
w'_{yi} = w_{yi} if document i also contains term x, or 0 otherwise
n = the number of documents in the collection

The term weight w_{xi} of term x in document i is computed using the standard $tf \times idf$ weighting formula. This algorithm computes the sum of maximum similarity values between each candidate translation of a term and the translations of all other terms in the query. They used one document as the window of term co-occurrence. It should be noted that the use of the Dice coefficient requires that the whole document be be used to collect co-occurrence statistics, and thus assumes that words that co-occur *anywhere* in the same document are correlated. This differs from the assumption underlying our technique, in which we assume that only words co-occur within a fixed window are likely to be contextually correlated. For each query term, the translation with the highest sum is selected as its translation. The results of their Indonesian–English and English–Indonesian CLIR experiments (58% and 74% of monolingual retrieval, respectively) demonstrated the effectiveness of this disambiguation technique. Maeda et al. [2000], working on Japanese–English CLIR, have used a search engine to collect the co-occurrence information between terms in web documents, and applied a modified Dice coefficient to calculate the mutual information between terms. They also used one document as the window of term co-occurrence. Gao et al. [2002] employed a similar approximate algorithm for choosing optimal translations in English–Chinese CLIR.

Federico and Bertoldi [2002] presented a statistical model for dictionary-based query translation. Translations are selected using a hidden Markov model (HMM) [Rabiner, 1990], which couples a translation lexicon with a bigram language model in the target language. Their query-translation model computes the probability of any query–translation (Italian–English) pair. This probability is modelled by a HMM in which the observable variable is the Italian query, and the hidden variable is its English translation. Their experimental evaluation of the CLIR model was performed on the Italian–English bilingual track data used in the CLEF 2000 and CLEF 2001 evaluations. Their best mean average precision value was 81% of monolingual.

A common feature of the selection algorithm used by all above approaches is that, while they use correlation between words, the word order has not been considered. Our idea is that the word order is important in query translation disambiguation. For example, ‘white house’ is not the same as ‘house in white’. The method we propose in this thesis attempts to address this issue.

Pirkola [1998] explored the effects of query structure and different translation dictionaries on the performance of CLIR. His study used a structuring method in which English translations that were derived from each Finnish query term were grouped into a set. A combination of a general machine readable dictionary and a domain specific medical lexicon was used in the Finnish–English query translation. Pirkola found that it is possible to solve the translation ambiguities and polysemy problems if queries are structured and if both general terminology and domain specific terminology are available in translation. The test collection consisted of the *Associated Press newswire*, *Federal Register*, and *Department of Energy abstracts* subsets of the TREC collection. This collection contained $514,825$ documents. Using the health related test queries selected from the TREC topics $1-300$, he was able to only achieve 77% of monolingual retrieval effectiveness.

One of the problems relating to research in this area is that most researchers have used different test collections and language pairs. They typically reported an improvement on their own baseline, however without a common standard it is very difficult to determine the contribution of each new technique. Thus the best we were able to do was to compare the results of the cross-lingual retrieval against those from the corresponding monolingual retrieval, in order to measure the effectiveness of the disambiguation methods. While this is probably the most reasonable way to compare the relative merits of each system, it should be noted that this is not completely objective. For example, we have observed that in general it is possible to obtain a greater improvement on a monolingual baseline when the underline monolingual system is poor. Additionally, the effect of using different language pairs in disambiguation experiments is not clear.

This problem of using non-standard collections was one of the prime motivating factors that lead to our participation in the NTCIR CLIR task. It not only allows to compare our results to those of others using the same data sets, but also permits subsequent researchers to compare their results with ours.

4.2 Improved Translation Disambiguation

We propose an improved technique that integrates a contextual window and a distance factor into a bigram Markov model to provide disambiguation for Chinese–English dictionary-based query translation. In contrast to previous techniques using statistics obtained from pre-defined training collections for translation disambiguation, our technique also uses web documents as a corpus to provide context for disambiguation.

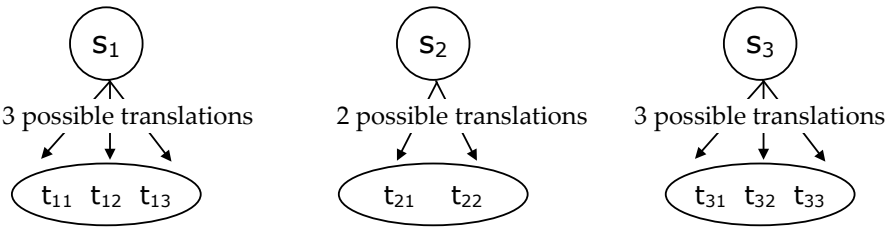


Figure 4.1: A graphical interpretation of translation ambiguity problem. Given a query with three query terms s_1, s_2, and s_3, for each query term, there are multiple translations provided by the dictionary.

$$\left\{\begin{array}{lll} t_{11}, t_{21}, t_{31} & t_{12}, t_{21}, t_{31} & t_{13}, t_{21}, t_{31} \\ t_{11}, t_{21}, t_{32} & t_{12}, t_{21}, t_{32} & t_{13}, t_{21}, t_{32} \\ t_{11}, t_{21}, t_{33} & t_{12}, t_{21}, t_{33} & t_{13}, t_{21}, t_{33} \\ t_{11}, t_{22}, t_{31} & t_{12}, t_{22}, t_{31} & t_{13}, t_{22}, t_{31} \\ t_{11}, t_{22}, t_{32} & t_{12}, t_{22}, t_{32} & t_{13}, t_{22}, t_{32} \\ t_{11}, t_{22}, t_{33} & t_{12}, t_{22}, t_{33} & t_{13}, t_{22}, t_{33} \end{array}\right\}$$

Figure 4.2: The set of all possible candidate translations of the given query s_1, s_2, s_3. Using a simple combination method, we generate a total of $3 \times 2 \times 3 = 18$ candidate translations.

4.2.1 Translation disambiguation using a Markov model

Given a query and a set of candidate translations, each candidate translation is a sequence of terms. Our idea is to estimate the likelihood of each sequence of terms using a probability model, and select the one with the maximum likelihood among all possible candidate translations as the most appropriate translation of the given query.

For example, given a query with three query terms s_1, s_2, s_3, we obtain a list of all possible translations for each of them via dictionary lookup. As shown in Figure 4.1, there are 3, 2, and 3 translations for the query term s_1, s_2, and s_3, respectively. Using a simple combination method, we generate a total of $3 \times 2 \times 3 = 18$ candidate translations of the given query (as shown in Figure 4.2). Each candidate translation is a sequence of terms t_{1x}, t_{2y}, t_{3z}, where $t_{1x} \in \{t_{11}, t_{12}, t_{13}\}$, $t_{2x} \in \{t_{21}, t_{22}\}$, and $t_{3x} \in \{t_{31}, t_{32}, t_{33}\}$. We then estimate the likelihood of each of these candidate translations and select the one with the maximum likelihood as the most appropriate translation of the query s_1, s_2, s_3. To estimate the likelihood of a sequence of terms, we use a probability model.

Markov chain model

In a n-gram Markov model, the probability of occurrence of a term is conditioned upon the prior occurrence of $n-1$ preceding terms. It is typically constructed from statistics obtained from a large corpus of text using the co-occurrences of terms in the corpus to determine term sequence probabilities. The Markov model is a stochastic finite-state automaton in which the states directly represent the observations. In the case of the language model this means that the states correspond to words. A Markov model is based on the Markov assumption that is the history of the current state is fully represented by the single previous state.

Computing the probability of a sentence $w_1, w_2, w_3,\ ...,\ w_n$ using a model to which this assumption is not applied would involve combining the probabilities for the separate words and their respective histories, as follows:

$$\begin{aligned} P(w_1, w_2, w_3,\ ...,\ w_n) &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ ...,\ P(w_n|w_1,\ ...,\ w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1,\ ...,\ w_{i-1}) \end{aligned}$$

It would be problematic to estimate the necessary probabilities from a corpus, since they concern long sequences that might not occur in the corpus even once. However, according to the Markov assumption, which is also called the limited horizon assumption, the immediate history of a word is representative for the whole of the word's history, leading to the following computation in case of a bigram model:

$$P(w_1, w_2, w_3,\ ...,\ w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\ ...\ P(w_n|w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_{i-1})$$

For larger values of n, larger amounts of history are incorporated into the individual states and the model becomes more accurate with respect to the corpus the n-gram model was based on.

Advanced model

Our disambiguation technique is based on a bigram Markov model [Markov, 1971]:

$$P(t_1, t_2, t_3,\ ...,\ t_n) = P(t_1) \prod_{a=2}^{n} P(t_a|t_{a-1})$$

To compute the probability of a sequence of words, we need to calculate the values of $P(t)$, the probability of word t, and $P(t|t')$, the probability of t in the context of t', as follows:

$$P(t) = \frac{f(t)}{N}\ ,\ P(t|t') = \frac{P_w(t, t')}{\sum_{t''} P_w(t'', t')}$$

where $f(t)$ is the collection frequency of term t, N is the number of terms in the document collection, and $P_w(t, t')$ is the probability of term t' occurring after term t within a window of size w.

The zero-frequency problem arises in the context of probabilistic language models, when the model encounters an event in a context in which it has not been seen before. Smoothing provides a way to estimate and assign the probability to that unseen event. Of the various smoothing methods available [Zhai and Lafferty, 2004], we use the following absolute discounting and interpolation formula, which applies the smoothing method proposed by Ney et al. [1994]. The idea of the absolute discounting method is to lower the probability of seen words by subtracting a constant from their counts. Federico and Bertoldi [2002] successfully used this method to compute the frequency of term t' and t within a text window of fixed size through an order-free bigram language model. In this method:

$$P(t|t') = \max\left\{\frac{f_w(t, t') - \beta}{N}, 0\right\} + \beta P(t)P(t')$$

where $f_w(t, t')$ is the frequency of term t' occurring after term t within a window size w, and the absolute discounting term β is calculated as follows:

$$\beta = \frac{n_1}{n_1 + 2n_2}$$

where n_k is the number of terms with collection frequency k.

We have observed that two words being in close proximity generally provides stronger correlation and produces more credible results for disambiguation of translation than does co-occurrence of two words in a large window. Gao et al. [2002] applied a decay factor to the mutual information calculation; their experiments showed that the decay factor can be used to discriminate strong and weak term correlation:

$$D(t, t') = e^{-\alpha(\mathrm{Dist}(t,t')-1)}$$

where $\mathrm{Dist}(t, t')$ is the average distance between t and t' in the document collection, α is the decay rate and determined empirically. We have incorporated this distance factor $D(t, t')$ into the probability calculation, to give:

$$P(t|t') = \left[\max\left\{\frac{f_w(t, t') - \beta}{N}, 0\right\} + \beta P(t)P(t')\right] \times D(t, t')$$

4.2.2 Utilization of web documents as a corpus

To reduce the ambiguity and errors introduced during query translation, various techniques utilizing statistics obtained from the test collections have been proposed [Adriani, 2000; Ballesteros and Croft, 1998; Federico and Bertoldi, 2002; Gao et al., 2002]. Using a pre-defined training corpus has two obvious drawbacks: first, a large training corpus is not always available and the construction of a large corpus requires costly human resources; second, any particular training corpus will always have limited coverage and will not satisfy the open domain problem. The growth of the web has made available vast written and spoken resources from almost all countries in the world. The web corpus can be used as practical resource for estimating the correlation of the translated terms. In this section, we investigate how to use the web to provide context for translation disambiguation.

Our idea is that the correct translations are generally semantically related and tend to co-occur more often in web documents than do incorrect translations. However, for the purpose of this research, it is impractical to use the entire web to collect corpus statistics. Our approach is to retrieve a set of web documents that contain the candidate translations of the query terms. We then use this subset of the web as a corpus to provide context for disambiguation of the particular query. We therefore need a mechanism to formulate queries which are most likely to retrieve the subset of pages that we are interested in. One important issue is that each original Chinese query term may have varying numbers of English translations. In order to prevent Chinese query terms with large numbers of English translations dominating the retrieval process, it is necessary to normalize the possible English translations in such a way that all possible English translations are treated equally for each Chinese query term. Not using normalization will lead to Chinese query terms with many possible translations getting higher scores than those having less possible translations.

Our procedure consists of three steps: bilingual dictionary lookup, structured query formulation, and web text processing.

Step 1 – Bilingual dictionary lookup

Suppose a query composed of a sequence of Chinese terms, we obtain a set of all possible English translations for each query term through bilingual dictionary lookup.

For example, consider a Chinese query “炭疽熱 | 細菌戰 | 恐怖 | 攻擊”[1] (anthrax | biological warfare | terror | attack) composed of four terms: “炭疽熱” (anthrax), “細菌戰 ” (biological war-

[1]This query (numbered 010) is taken from the NTCIR−5 collection, in which the Chinese query terms are pre-segmented.

Chinese query terms	English translations
炭疽熱	anthrax
細菌戰	biological warfare | germ warfare
恐怖	terror | consternation | funk | monstrousness
攻擊	attack | accuse | aggress | bombard | impugnment | onfall | pounce | scuff

Table 4.1: English translations obtained for "炭疽熱 | 細菌戰 | 恐怖 | 攻擊" through a bilingual dictionary lookup.

fare), "恐怖" (terror), and "攻擊" (attack). For each query term, we obtain all English translation equivalents provided by the bilingual dictionaries, as shown in Table 4.1.

Step 2 – Structured query formulation

We then transform the Chinese query into a structured English query using the English translation equivalents obtained. The structured query is formulated according to the following rules:

1. The translation sets of all query terms are combined with the logical operator AND;
2. The candidate translations of a query term are enclosed in the parentheses and combined with the logical operator OR (the parentheses are used to indicate the order in which we want the search engine to interpret the Boolean operators.);
3. Phrases are enclosed in quotation marks as units.

Extending the example used in Table 4.1 , the Chinese query "炭疽熱 | 細菌戰 | 恐怖 | 攻擊" is transformed into a structured English query as follows:

(anthrax) AND
("biological warfare" OR "germ warfare") AND
(terror OR consternation OR funk OR monstrousness) AND
(attack OR accuse OR aggress OR bombard OR impugnment OR onfall OR pounce OR scuff)

Step 3 – Web document processing

Boolean algebra queries are supported by most search engines. We use several search engines to fetch up to 300 top-ranked documents using the structured English query generated in the previous step. The retrieved query-biased web documents are then filtered to remove HTML tags and metadata, leaving only the web text as the corpus to provide context for disambiguation.

4.3 Disambiguation Experiments

In our disambiguation approach, we integrated window size and a distance factor into a bigram Markov model for disambiguation. In addition, we used statistics obtained from the web documents as a corpus, rather than a specific test collection, to provide context for disambiguation.

In this section, we wish to explore issues related to translation disambiguation only: first, we investigate the effect of context window size and distance factor on disambiguation performance; second, we are interested to see how well our disambiguation technique performs when the web is used to provide context; and third, we compare our disambiguation technique to other techniques, that have different underlying principles, on the same data sets.

We ran a set of experiments using the data sets of NTCIR 4 and 5 as described in Section 2.4.3. A sample Chinese topic is shown in Figure 2.9. We choose to use the titles of the Chinese topics only as queries to retrieve English documents for two reasons: first, ambiguity problem is often resolved implicitly when queries are long enough (the additional words provide sufficient context to resolve confusion) but is still a critical problem when queries are short; second, web queries are often short, and the average length of the titles approximates that of web queries. We use the *rigid relevance judgements* provide by NTCIR in our results.

The full details of our CLIR experimental setup are given in Chapter 5.

4.3.1 Effect of window size and distance factor on translation disambiguation

Using the NTCIR−4 English collection and the titles of the associated 58 Chinese topics, we have gathered in Table 4.2 a side-by-side comparison of the average precision values for a total of 36 runs using the combinations of the different values of the context window size w and the decay rate α (as described in Section 4.2.1).

Our use of window size is based on the assumption, that when two words co-occur in the same query, they are likely to co-occur in the same text windows in the documents. Conversely, two words that do not co-occur in the same text window are not likely to show up in the same query. Therefore, the correct translation of the given query term is not only determined by the immediately adjacent words, but also by the words that co-occur in the same text window. Generally, the larger the window size, the more contextual information will be provided; and consequently more computational cost will be involved. Our experimental results showed that that variation in window size used to collect word association information has a small effect on the outcome, with $w = 4$ producing the slightly better results of Chinese–English CLIR.

Gao et al. [2002] reported that setting $\alpha = 0.8$ gave the best results when combined with their

w	$D(t,t') = e^{-\alpha(\mathrm{Dist}(t,t')-1)}$, where $\alpha =$					
	0.0	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
2	0.2112	0.2112	0.2112	0.2112	0.2112	0.2112
4	**0.2166**	0.2166	0.2166	0.2166	0.2166	0.2166
6	0.2152	0.2152	0.2152	0.2152	0.2152	0.2152
8	0.2152	0.2152	0.2152	0.2152	0.2152	0.2152
10	0.2161	0.2161	0.2161	0.2161	0.2161	0.2161
14	0.2160	0.2160	0.2160	0.2160	0.2160	0.2160

Table 4.2: Effect of window size and distance factor on translation disambiguation in Chinese–English CLIR, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. w is the window size and α is the decay rate. $w = 4$ produced the slightly better result 0.2166. There is no difference for values of α between $[0, 1]$. This set of experiments used the Zettair IR system.)

mutual information model. However, our experimental results showed that there was no difference for values of α between 0.0 and 1.0. It can be seen that our test collection is insensitive to this parameter. The explanation seems to be that the Markov model is a superior technique where sequence data is involved, and is not improved by the addition of a decaying distance factor.

4.3.2 Disambiguation using web documents versus test collections

We next investigate how well our translation disambiguation technique performs when the web is used to provide context for disambiguation. We use the English document collection from the NTCIR−5 CLIR task and the associated 49 Chinese topics (see Section 2.4.3). We employ the NTCIR−5 English document collection and the English web documents extracted by a search engine as a corpus to disambiguate dictionary translation, respectively. This set of experiments used the Lemur IR system[2] developed by the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

Our retrieval experiments consist of four runs. In T–*runs*, we have used the titles of the Chinese topics (as shown in Figure 2.9) as queries, and in D–*runs* the descriptions are used as queries to retrieve the documents from the English document collection. We established a mono-lingual reference (T−*mono* and D−*mono*) by which we can measure our CLIR results. We then tested the translation disambiguation using the test collection (see T−*collection* and D−*collection*) and

[2] `http://www.lemurproject.org/`

RunID	Average precision	Recall	P@10
T−*mono*	0.4564	0.9444	0.5286
T−*collection*	**0.3702**	0.8660	0.4388
T−*web*	0.3428	0.8661	0.4143
D−*mono*	0.4391	0.9362	0.5429
D−*collection*	0.3917	0.8674	0.4959
D−*web*	**0.4042**	0.8779	0.4857

Table 4.3: Chinese–English translation disambiguation using the extracted English web documents and the NTCIR−5 English test collection, respectively. We note that the average precision values of T−*collection (0.3702) and* D−*web (0.4042) runs were the best results amongst all participants in the NTCIR−5 Chinese–English CLIR task. (Average precision, recall, and P@10 for the titles* T−*runs and descriptions* D−*runs of NTCIR−5 Chinese topics. The queries are translated using the combination of translation disambiguation and OOV term translation (see Section 3.2) techniques. This set of experiments used the Lemur IR system.)*

the English web documents extracted by a search engine, (see T−*web* and D−*web*), respectively.

The results of our experiments are shown in Table 4.3. We note that the average precision values of T−*collection* and D−*web* runs were the best results amongst all participants in the NTCIR−5 Chinese–English CLIR task. The average precision values for those two runs were 0.3702 and 0.4042, respectively representing 67.3% and 92% of monolingual retrieval effectiveness. We used the Wilcoxon signed-rank test, which is described in more detail in Section 2.4.1, to test the statistical significance of our results. It showed that using our disambiguation technique, we were able to use web data for translation disambiguation, rather than a specific collection, with no significant loss of effectiveness. When using the web, it is possible to achieve effectiveness comparable to that obtained with a pre-defined training corpus.

4.3.3 Comparison of different disambiguation techniques

The disambiguation phase is crucial to the translation result. Interestingly, researchers working on CLIR between Asian languages and English have used different disambiguation techniques utilizing statistics obtained from the test collection, all seemingly with good results. Term similarity approaches [Adriani, 2000; Maeda et al., 2000] use both tf and idf, and select the best translation of each query term from multiple candidates by comparing their statistical associations with the candidate translations of all other query terms within each document. Term co-occurrence

	RUN−*ts*	RUN−*tc*	RUN−*markov*
Model	Dice similarity coefficient	Mutual information	Bigram Markov model
Co-occurrence frequency	×	√	√
Term distance	×	√	×
Term order	×	×	√
Term weighting	$tf \times idf$	tf only	tf only
Window size	A document	A sentence	4 terms
Translation selection	The translations of all other query terms are considered.		Only the translations of the immediately preceding query term are considered.

Table 4.4: A comparison of various translation disambiguation techniques. (The term similarity technique in RUN−*ts, the term co-occurrence technique in* RUN−*tc, our Markov model technique in* RUN−*markov.)*

approaches [Ballesteros and Croft, 1998; Gao et al., 2002; Kwok, 2000; Jang et al., 1999] make use of tf and the frequency of terms co-occurring within a window size of w in the collection. In addition, a distance factor is incorporated to discriminate strong and weak term correlation.

As these previous experiments were generally carried out on different test collections, it is unclear whether any particular approach is superior. In this section, by using the same data sets, we compare our disambiguation technique (as explained in Section 4.2.1) to others with different underlying principles — term similarity measure [Adriani, 2000] and term co-occurrence [Gao et al., 2002] — in Chinese–English CLIR; and thus understand the merits of the different disambiguation techniques. The formulae and the parameters used by Adriani [2000] and Gao et al. [2002] were described in Section 4.1. The characteristics of the three different disambiguation approaches are tabulated in Table 4.4.

Comparison results and discussions

We ran a set of experiments using the data sets of NTCIR 4 and 5 (as described in Section 2.4.3) to compare the three techniques in Chinese–English CLIR. Our experiments consist of four runs: a mono-lingual reference in RUN−*mono*, the disambiguation techniques using term similarity in RUN−*ts*, term co-occurrence in RUN−*tc*, and our approach based on a Markov model in RUN−*markov*. Each of the techniques uses different models, formulae and parameters; nonethe-

	NTCIR−5 data set			
	RUN−*mono*	RUN−*ts*	RUN−*tc*	RUN−*markov*
Average precision	0.3629	0.3117	0.3158	0.3117
% Monolingual	—	85.9	87.0	85.9
Recall	0.8607	0.7827	0.7984	0.7903
P@10	0.4755	0.4082	0.4143	0.3796
	NTCIR−4 data set			
	RUN−*mono*	RUN−*ts*	RUN−*tc*	RUN−*markov*
Average precision	0.2423	0.1985	0.1981	0.200
% Monolingual	—	81.9	81.8	82.5
Recall	0.6877	0.6201	0.6518	0.6457
P@10	0.4034	0.3310	0.3259	0.3345

Table 4.5: Chinese–English CLIR results using different disambiguation techniques — term similarity, term co-occurrence, and a bigram Markov model. Average precision, recall, and P@10 for the titles of NTCIR 4 and 5 Chinese topics. (Data consists of the NTCIR 4 and 5 English collections. Neither OOV term translation nor post-translation query expansion is applied in the query translation process. This set of experiments used the Lemur IR system.)

less each achieved comparable results across multiple data sets, as shown in Table 4.5.

We carried out a Wilcoxon signed-rank test (previously described in 2.4.1) to analyze the statistical significance of our results. These results suggest that there is no significant difference in the overall effectiveness on the data sets tested. However, analysis of the individual queries reveals a different story. There are 58 Chinese topics in NTCIR−4 and 49 Chinese topics in NTCIR−5. However, not every query contains ambiguous Chinese query terms and such queries are not affected by the ambiguity problem. Of the 93 NTCIR 4 and 5 queries with ambiguity (at least one query term has multiple translations), the three different disambiguation techniques produced the same translation in only 19% of the queries (10 out 50 NTCIR−4 queries and 8 out of 43 NTCIR−5 queries), and manual inspection showed that these were indeed correct translations. In the other 81% of queries with differing translations, there was often significant difference in the average precision results for each of the approaches. An indicative sample of the differing results for the 50 NTCIR−4 ambiguous queries is shown in Figure 4.3.

In some cases there were only minor differences in average precision. For example, in query 040, the Chinese term “问题” was variously translated as “problem” and “issue”. This term has

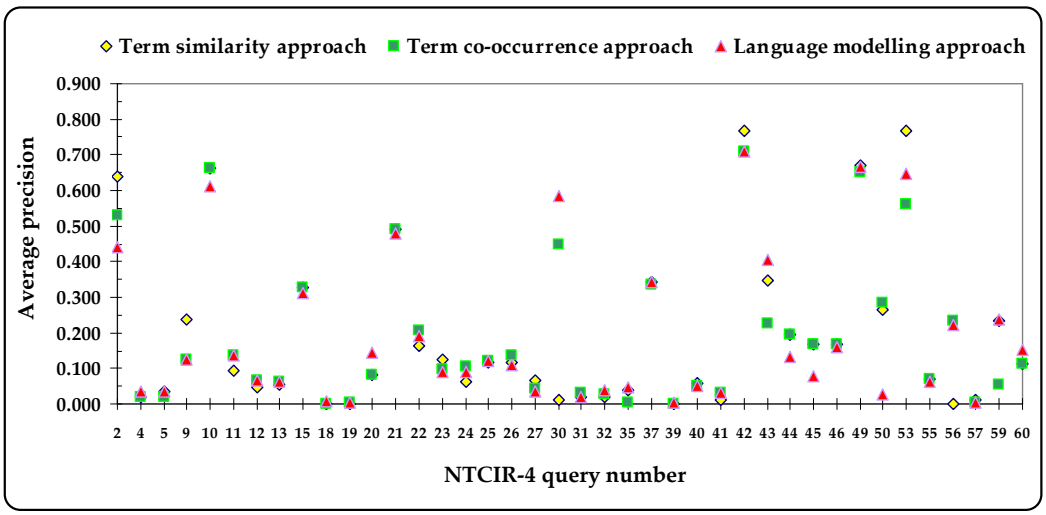


Figure 4.3: An example of different average precision values obtained using different disambiguation techniques — term similarity, term co-occurrence, and a bigram Markov model — for individual queries. (Data consists of the NTCIR−4 English collections and the titles of the NTCIR−4 Chinese topics.)

a low weight and the other more significant terms — "足球" (soccer), "世界杯" (world cup), and "门票" (ticket) — were correctly translated. However in other cases, one or more technique has effectively failed, as a result of selecting a totally inappropriate translation. For example, in query 030 "动物复制技术", the term similarity approach translated the query as "animal copy skill", whereas the other methods correctly translated the query as "animal clone technology".

This research has shown that, superficially, all these translation disambiguation methods are comparable (no significant difference) when averaged across a query set. However, at the individual query level, each of the techniques frequently produce differing results. This means that it may be possible to develop a new approach that combines the best elements of these current methods. At the very least, it should be possible to use a "combination of evidence" approach [Smets, 1990] to improve the overall translation quality. Further, when all methods produce the same translation, we can have a higher degree of confidence in the correctness, and differing translations could act as a trigger for further analysis.

4.4 Summary

The translation ambiguity problem is particularly severe in view of the observed tendency of web users to enter short queries; in general it is difficult for even a human to reliably determine the intended meaning from the available context. In this chapter, we proposed a disambiguation technique for dictionary-based query translation. This simple technique has proved to be extremely robust and successful.

Our disambiguation technique is based on a Markov model [Markov, 1971]; such models have been used widely for probabilistic modelling of sequence data. Two words being in close proximity generally provides stronger correlation and produces more credible results for disambiguation than does co-occurrence of two words in a large window. We therefore investigated the effects of distance factor and window size when using a Markov model to provide disambiguation. Contrary to what has been noted when using mutual information techniques to provide disambiguation, we observed that using a window distance factor has no benefit when combined with a Markov model. This showed that the Markov model is a superior technique where sequence data is involved, and is not significantly improved by the addition of a distance factor, thus potentially reducing the computational cost.

In previous work, researchers have relied on the test collection corpus to perform translation disambiguation in CLIR. Using a pre-defined training corpus for disambiguation has limitations in both availability and coverage. Since the web consists of documents in various domains or genres, we investigated how to utilize the web documents as a corpus to provide context for translation disambiguation. Our experimental results showed that when using the web, it is possible to achieve effectiveness comparable to that obtained with a pre-defined training corpus.

We also explored other alternatives for translation disambiguation. To understand the relative merits of the different disambiguation techniques, we compared our technique to other approaches on the same Chinese–English data sets. Our results showed that, despite the different underlying models and formulae used, the aggregated results are comparable. However, there is wide variation in the translation of individual queries, suggesting that there is scope for further improvement.

Chapter 5

Query Translation Experiments

The aim of this thesis is to develop and test new techniques for Chinese–English query translation, and thus improve the effectiveness of Chinese–English CLIR. Query translation includes several separate elements. In the previous chapters, we discussed the individual techniques — OOV term translation and translation disambiguation — that we have developed as part of our dictionary-based query translation process. In this chapter we explore the combinations of these techniques and investigate the improvement contributed by each component. Additionally, most CLIR systems commonly apply some form of query expansion technique, as this typically provides improvement in CLIR effectiveness. Although not the main focus of this thesis, we nonetheless investigate the effect of post-translation query expansion on Chinese–English CLIR.

This set of Chinese–English CLIR experiments used the Zettair search engine developed by the Search Engine Group[1] at RMIT University. We used the test collection from the NTCIR−4 workshop Chinese–English CLIR task. As previously described in Section 2.4.3, this test collection consists of an English document set (347,376 news articles between 1998 to 1999), 58 Chinese topics, and relevance judgments for each search topic.

In Section 5.1, we present our two-stage procedure of selecting additional query terms for post-translation query expansion based on term weighting and word association information. In Section 5.2, we explain our dictionary-based Chinese–English query translation process step-by-step. In Section 5.3, we describe how we set up our experiments. In Section 5.4, we show our experimental results along with the conclusions.

[1] `www.seg.rmit.edu.au`

5.1 Post-translation Query Expansion

Chinese web queries may contain some words that are not included in the segmentation dictionary, and thus these types of terms may not be correctly translated into the corresponding English equivalents when using a bilingual dictionary. In order to ameliorate these translation errors, query expansion which is widely used in monolingual IR, may be employed to overcome these problems.

In this section, we first summarize previous work in query expansion. Following that we present our two-stage procedure of selecting additional query terms for post-translation query expansion based on term weighting and word association information.

5.1.1 Related work

Query expansion by pseudo-relevance feedback is a well-established procedure in both monolingual and cross-lingual IR, potentially providing some improvement in retrieval effectiveness. Expansion in the context of CLIR is particularly interesting as it presents multiple opportunities for improving retrieval effectiveness. The pseudo-relevance feedback process can take place prior to translation, afterwards, or at both stages. In CLIR, pre-translation query expansion means using a separate collection in the query language for pre-translation retrieval in order to expand the query with highly associated terms. These terms may help focus on the query topic and bring more translated terms that together are useful for disambiguating the translation. In a typical form of post-translation query expansion, t terms from the top d documents retrieved using the translated query, are selected and added to the translated query.

Ballesteros and Croft [1997] evaluated pre- and post-translation query expansion in a Spanish–English CLIR task and found that combining pre- and post-translation query expansion improved both precision and recall with pre-translation expansion improving both precision and recall, and post-translation expansion enhancing precision. In a subsequent study, Ballesteros and Croft [1998] examined the use of co-occurrence statistics in parallel corpora to select translations from a machine-readable dictionary. This technique boosted bilingual performance from 68% to 88% of a monolingual baseline. Here they suggested that post-translation expansion helps remove errors due to incorrect translations. McNamee and Mayfield [2002]'s dictionary ablation experiments on the effect of translation resource size and pre- and post-translation query expansion effectiveness demonstrated the key and dominant role of pre-translation expansion in providing translatable terms. They observed that post-translation expansion can provide little improvement if too few terms are translated. They explored the five language pairs Dutch, French, German, Italian, and

Spanish, to English in their CLIR experiments.

Interestingly, the benefits of query expansion for Chinese–English and English–Chinese CLIR are unclear. Gey and Chen [2000] wrote an overview of the TREC-9 CLIR track, which focused on using English queries to search a Chinese news collection. Their summary of work by several top-scoring track participants reveals a disconcerting lack of consistency as to the merits of query expansion methods:

- 10% improvement in average precision with either pre-translation or post-translation expansion (reported by Xu and Weischedel [2000]);
- Pre-translation query expansion did not improve the overall precision (reported by Gao et al. [2000]);
- The best run of Fudan University did not use post-translation expansion (reported by Wu et al. [2000]);
- The best cross-language run of Chinese University of HongKong did not use post-translation expansion (reported by Jin and Wong [2000]);
- The best run of Queens College used both pre- and post-translation expansion (reported by Kwok et al. [2000]);
- Post-translation query expansion yielded very little improvement (reported by Allan et al. [2000]).

With inconsistent results like these, it is impossible to ascertain which techniques do and do not work. Using the data collection from NTCIR−4 Chinese–English CLIR task, Levow [2004] claimed a dramatic improvements from post-translation query expansion. However, in comparison to other participants of the same task, their overall results are mediocre. They reported that the increases due solely to the pre-translation query expansion were much smaller. Similarly, Kwok et al. [2004] showed that pre-translation query expansion degraded the average precision. In the contrast, using only post-translation query expansion provided the best Chinese–English CLIR results. However, the improvements were not statistically significant at a 95% confidence level.

In this thesis, we use NTCIR−4 test collection to test our post-translation query expansion technique, and also as a basis for comparison to other participants in the same CLIR task. Additionally, we investigate the impact of post-translation query expansion on Chinese–English CLIR retrieval effectiveness, in particular, when the Chinese queries are imperfectly translated (the

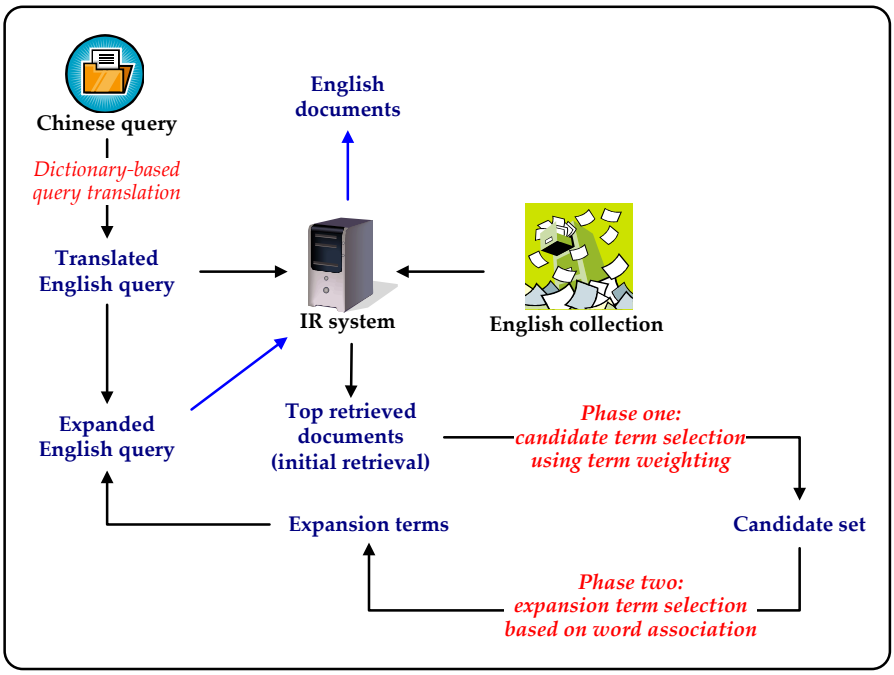


Figure 5.1: Flow chart of the post-translation query expansion process.

translated queries are potentially faulty). We note that query expansion involves selection of parameter values that are not necessarily consistent from one collection to another [Billerbeck and Zobel, 2004].

We start by describing our automatic procedure for post-translation query expansion. After that, we investigate the interaction of the parameters used.

5.1.2 Two-phase post-translation query expansion process

The motivation of our post-translation query expansion technique is that, not only should terms be "important" in terms of having a high weight, but they should be related to the original query. We propose a two-phase process of an automatic feedback query expansion: candidate term selection and expansion term selection. As shown in Figure 5.1, during the first phase, we select a set of candidate terms from the top d retrieved documents using standard term weighting metrics. In the second phase, we apply a statistical measure of word association to select the top t expansion terms from the initial candidate term set, and add them to the original query. In a post-translation query expansion procedure, the original query refers to the translated query.

In the following subsections, we first describe how to determine the importance of a candidate

t		d					
		5	10	20	30	40	50
0	—	0.2166					
5	tf	0.2157	0.2218	0.2259	0.2280	0.2280	0.2104
	$tf \times idf$	0.2060	0.2286	0.2211	0.2224	0.2254	0.2239
10	tf	0.2158	0.2258	0.2175	0.2102	0.2022	0.2036
	$tf \times idf$	0.2098	0.2202	0.2106	0.2041	0.2016	0.1933

Table 5.1: Post-translation query expansion using tf and $tf \times idf$, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. d is the number of the top ranked documents returned and t is the number of the expansion terms added. It can be seen that there is not a great difference in the results, but in all cases using tf to select the top terms was more effective than using $tf \times idf$.)

term in phase one. We then explain how to calculate the degree of word association between a candidate term and a set of original query terms in phase two.

Phase 1 – Candidate term selection

In the candidate term selection phase, for each translated query, we select a set of candidate terms based on their importance in the top d retrieved documents. We experimented with two term weighting approaches, tf and $tf \times idf$, to the selection of the initial candidate set. The value of tf is the frequency with which the term occurs in the top d retrieved documents. The value of idf is calculated as $\log(N/d_f)$, where d_f is the document frequency of the term and N is the total number of documents in the document collection. English stop words were removed from the retrieved documents prior to term selection.

To provide a baseline for our query expansion experiments using word association information, we experimented with adding either either 5 or 10 top-ranked terms from the top-ranked 5, 10, 20, 30, 40, and 50 documents, without the additional second phase. The results of these experiments are shown in Table 5.1. It can be seen that there is not a great difference in the results, but in all cases using tf to select the top terms was more effective than using $tf \times idf$. It can also be seen that using more documents provided a slight improvement up to 30 documents, but no improvement after that.

Phase 2 – Expansion term selection

As explained above, the second phase involves selecting the top t expansion terms from the set of candidate terms that have the highest degree of word association with all terms in the translated query. We use a well-known statistical method, the *Mutual information* (MI) measure [Church and Hanks, 1990], to estimate the strength of association between two terms.

The MI measure has been applied successfully to various IR problems. Pantel et al. [2005] used the MI to match database columns containing similar information. Wang and Lochovsky [2004] proposed a novel feature selection method, in which the MI is used to measure the relationships among the features, for text categorization. Fleuret [2004] also developed a very fast feature selection technique based on the conditional MI. Jang et al. [1999] suggested a disambiguation method using MI statistics obtained from the target collection in Korean-English CLIR. They also used the MI to assign a weight to translation English query terms. Zhang et al. [2000] introduced a statistical method, that is based on mutual information and context dependency, to extract Chinese compound words from a very large corpus. It should be noted that other statistical methods of association such as chi-square tests, log-likelihood ratios, z-scores or t-scores could also be used to measure the term correlation. In this thesis, however, we focus on the MI only as a proof of principle that word association information in query expansion is likely to provide any benefit.

To measure the MI between a given term x and a term s within a window size of w, we used:

$$\mathrm{MI}(x, s) = \log\left(\frac{f_w(x, s)}{f_x f_s} + 1\right)$$

where $f_w(x, s)$ is the frequency with which x and s co-occur within a window size of w in the document collection; f_x is the collection frequency of x; and f_s is the collection frequency of s. Addition of 1 to the frequency ratio means that a zero co-occurrence frequency corresponds to zero mutual information. The MI compares the probability of x and s together (the joint probability) with the probability of x and s independently (chance). In order to select the best t expansion terms from the candidate set obtained in phase one, we need to calculate the mutual information between a term and a term set. The MI of a term x and a set S of terms is the sum of x with every term in the set S, as follows:

$$\mathrm{MI}(x, S) = \sum_{s \in S} \mathrm{MI}(x, s)$$

5.1.3 Parameter selection

The fact that we are using a two-phase process requires that we decide how many candidate terms to collect in the first phase, which we call the candidate set c, to consider in the second phase. We express this as a proportion of the number of expansion terms added. For example, if we add t expansion terms to the original query, we might collect an initial candidate set of $c = 2t$. In addition to selecting the number of terms t and documents d that participate in query expansion, use of the MI also requires selection of a window size w in which mutual information statistics are collected. Our post-translation query expansion thus involves four parameters:

d: the number of the top ranked documents returned

t: the number of the expansion terms added

w: the window size used in mutual information

c: the size of the candidate set

There are many possible combinations of these parameters. We investigate the interaction of these parameters, as follows.

Effect of adding documents. As neither tf nor $tf \times idf$ was clearly superior (as described in Section 5.1.2), we experimented with using both of these to select terms for the next phase. From further experiments (not presented here) we determined that choosing $w = 4$, 16, or 20 and $t = 5$ or 10 gave slightly better results. Results are shown in Table 5.2. From this table we can observe slight improvements up to 20 documents, and a decline after that. We also note the adding 5 terms is always superior to using 10 terms, something we explore further below. Finally, using tf together with MI is more effective than using $tf \times idf$ with MI.

Effect of adding terms. Although our previous experiments (Table 5.2) suggest that $t = 5$ may provide the best results, we experimented with adding larger numbers of terms to see if this had any effect. As tf had proven superior to $tf \times idf$, we persevered only with tf. The results of these experiments are shown in Table 5.3, and confirm that $t = 5$ produces the best results.

Effect of window size. Table 5.4 shows the effect of window size on query expansion using mutual information. Again we have chosen other parameter values that appear to give optimal results. From the results in Table 5.4, we can see that there is no consistent trend, although $w = 16$ gives the best results when $d = 20$. Once again $t = 5$ performs best.

d		w = 4		w = 16		w = 20	
		$t = 5$	$t = 10$	$t = 5$	$t = 10$	$t = 5$	$t = 10$
5	tf	**0.2264**	0.1965	0.2249	0.1983	0.2246	0.2022
	$tf \times idf$	**0.1947**	0.1947	0.1815	0.0.1936	0.1848	0.1906
10	tf	**0.2314**	0.2154	0.2272	0.2127	0.2244	0.2210
	$tf \times idf$	**0.2185**	0.2119	0.2114	0.2101	0.2111	0.2073
20	tf	**0.2294**	0.2051	0.2386	0.2075	0.2341	0.2097
	$tf \times idf$	**0.2171**	0.2117	0.2049	0.1964	0.2068	0.2016
30	tf	**0.2254**	0.2090	0.2247	0.2084	0.2271	0.2109
	$tf \times idf$	**0.2119**	0.2054	0.2092	0.1982	0.2105	0.1985
40	tf	**0.2153**	0.2055	0.2168	0.2043	0.2185	0.2049
	$tf \times idf$	**0.2163**	0.1929	0.2193	0.1941	0.2201	0.1962

Table 5.2: Effect of the number of the top-ranked documents returned on post-translation query expansion, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. d is the number of the top ranked documents returned, w is the window size used in MI, and t is the number of the expansion terms added. Slight improvements up to $d = 20$, and a decline when $d > 20$. Using $t = 5$ is always superior to using $t = 10$.)

w	t = 5	t = 10	t = 20	t = 30
4	**0.2294**	0.2051	0.2027	0.1838
8	**0.2263**	0.2019	0.2057	0.1875
12	**0.2279**	0.2065	0.2052	0.1862
16	**0.2386**	0.2075	0.2054	0.1903

Table 5.3: Effect of number of expansion terms added on post-translation query expansion, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. w is the window size used in MI and t is the number of the expansion terms added. $t = 5$ always gives the best results.)

Effect of candidate set size. A final issue is the number of terms collected in the first phase for consideration in the second phase. In the above experiments we used a candidate set that was twice the number of terms ultimately required, namely $c = 2 \times t$. We wondered if collecting more

t	w						
	4	8	12	16	18	20	22
5	0.2294	0.2263	0.2279	**0.2386**	0.2293	0.2241	0.2242
10	0.2051	0.2019	0.2065	0.2075	0.1974	0.2097	0.2074

Table 5.4: Effect of window size on post-translation query expansion, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. w is the window size used in MI and t is the number of the expansion terms added. No consistent trend, although $w = 16$ gives the best results 0.2386 when $d = 20$. Once again $t = 5$ performs best.)

c	d					
	5		10		20	
	$t = 5$	$t = 10$	$t = 5$	$t = 10$	$t = 5$	$t = 10$
$t \times 2$	**0.2249**	0.1983	0.2272	0.2127	0.2386	0.2075
$t \times 3$	0.1941	0.1974	0.2188	0.2151	0.2245	0.2012
$t \times 4$	0.1877	0.1957	0.2114	0.2018	0.2052	0.2042

Table 5.5: Effect of candidate set size on post-translation query expansion, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. c is the size of the candidate set, w is the window size used in MI, and t is the number of the expansion terms added. $t = 5$ and $c = 2 \times t = 10$ give the best results 0.2249.)

terms in the first phase might improve results, so we experimented with using larger candidate sets, namely $c = t \times 3$ and $c = t \times 4$. However as can be seen in Table 5.5, this only led to a deterioration in performance.

5.1.4 Significance of word association in query expansion

After testing a large number of combinations, as outlined above, we selected the best results from query expansion using only tf to compare with the best results provided by query expansion using tf with MI. While this represents tuning, it allowed us to test whether using word association information in query expansion is likely to provide any benefit.

We used the the Wilcoxon signed-rank test (described in more detail in Section 2.4.1) to test the statistical significance of our results. Our baseline title run T−*do* using disambiguation and OOV translation (see Section 5.3, Table 5.7) achieved an average precision of 0.2166. As shown in Table 5.6, using query expansion based on tf with $t = 5$ and $d = 30$, we achieved 0.2280 which

	w	d			
		5	10	20	30
tf with Mutual Information	4	0.2264	0.2314	0.2294	0.2254
	16	0.2249	0.2272	**0.2386**	0.2247
	20	0.2246	0.2244	0.2341	0.2271
tf	–	0.2157	0.2218	0.2259	0.2280

Table 5.6: Effect of using tf with and without MI on post-translation query expansion, average precision for the titles of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection. w is the window size used in MI and t is the number of the expansion terms added. Using tf and MI, $d = 20$, $t = 5$, $c = 2t$, and $w = 16$, give the best results 0.2386.)

represents a 5% improvement; using tf and MI, with $d = 20$, $t = 5$, $c = 2t$, and $w = 16$, we achieved 0.2386, which represents a 10% improvement. However neither of these improvements is statistically significant.

5.2 Chinese–English Query Translation Process

In our Chinese–English CLIR experiments, we translate the titles and descriptions of the NTCIR−4 Chinese topics into English using the combinations of the techniques we developed, and then use the translated English queries to retrieve the documents from the English document collection. We note that the average length of the titles is around 3.3 terms, which approximates the average length of web queries. Our dictionary-based Chinese–English query translation process is shown in Figure 5.2.

English stop words were removed from the English document collection. We use a stop list that contains 477 entries and the Porter stemmer [Porter, 1980] to reduce words to stems. The Chinese queries were processed as follows:

Pre-processing. In NTCIR−4, the title of each Chinese topic is presented as a list of comma-separated Chinese text (as shown in Figure 2.9). Our assumption is that each of these text strings is either a phrase or a word. If a Chinese string cannot be found in the translation dictionaries, we treat it as a potential Chinese OOV term.

Step 1 – Chinese OOV term translation. Using the potential Chinese OOV terms obtained as queries, we apply our Chinese OOV term translation technique (described in Chapter 3). We

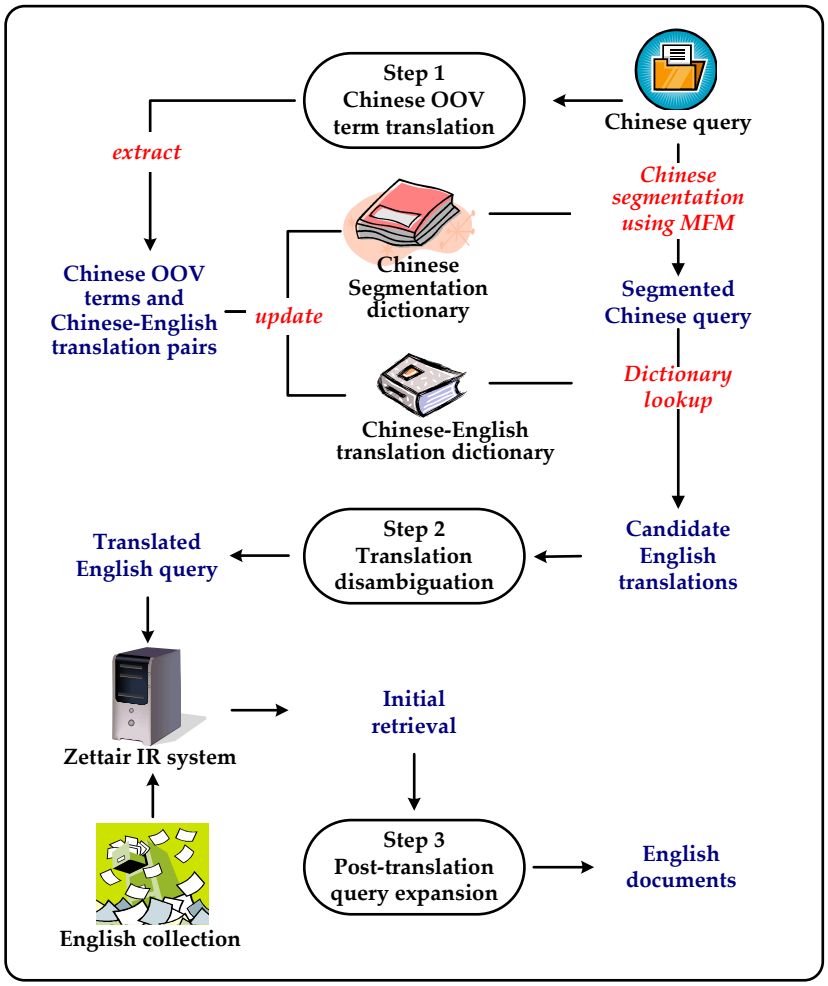


Figure 5.2: Flow chart of dictionary-based Chinese-English query translation process.

detect Chinese OOV terms using a segmentation-free process and add new terms into both of the Chinese segmentation dictionary (see Section 2.3.3) and the BIG5–English translation dictionary.

We then use the updated Chinese segmentation dictionary to segment the queries and replace each Chinese query term using all possible English translations obtained from the updated BIG5–English translation dictionary. A dictionary-based word segmentation method with the maximum forward matching (described in Section 2.3.3) is adopted to segment the Chinese queries.

Step 2 – Translation disambiguation. Once candidate query term translations are collected, we apply our disambiguation technique (described in Chapter 4) to determine the most appropriate English translation for each Chinese query term.

Step 3 – Post-translation query expansion. Based on the results in Figure 5.6, for each translated query, we first select 10 candidate terms based on the tf weighting in the top 20 retrieved documents; and then apply a mutual information (using a window size of 16) procedure to select the best 5 expansion terms and add to the translated query.

5.3 Experimental Design

Our CLIR experiments consist of sixteen runs. In T−*runs*, we have used the titles of the Chinese topics as queries, and in D−*runs* the description fields are used as queries to retrieve the documents from the English document collection. The relevance judgements provided by NTCIR are at two levels — strictly relevant documents known as *rigid relevance*, and documents that are likely to be relevant, known as *relaxed relevance*. We use only *rigid relevance* in our results.

To provide a baseline for our CLIR results, we used BabelFish to "manually" translate each Chinese query. The retrieval results are shown as runs T−$BabelFish$ and D−$BabelFish$. Kraaij [2001] showed successful use of the BabelFish translation service based on Systran. We established a mono-lingual reference (T−*mono* and D−*mono*) by which we can measure our CLIR results. If the Chinese queries were translated perfectly, we would expect to achieve the same retrieval effectiveness as mono-lingual retrieval. We then tested disambiguation and OOV translation. After each stage, we have also tested query expansion. These experiments allow us to separately gauge the improvement contributed by each of our techniques. A brief description of the runs is shown in Table 5.7.

5.4 Results and Discussion

In this section we explore the combinations of the three techniques — OOV term translation (in Chapter 3), translation disambiguation (in Chapter 4), and post-translation query expansion (in Section 5.2) — that we have developed, and investigate the improvement contributed by each component in Chinese–English CLIR. The results of these experiments are shown in Table 5.8.

5.4.1 BabelFish and disambiguation

In previous work [Zhang and Vines, 2003] using the NTCIR−3 query set, we found that disambiguation alone was always more effective than the BabelFish baseline. This was not the case with the NTCIR−4 query set. Although the average precision for the description runs was slightly higher using the disambiguation technique, the average precision for the *title* runs was lower than

RunID	Translation Disambiguation *(d)*	OOV Translation *(o)*	Query Expansion *(q)*
T–*BabelFish*	×	×	×
T–*BabelFish+q*	×	×	√
T–*d*	√	×	×
T–*dq*	√	×	√
T–*do*	√	√	×
T–*doq*	√	√	√
D–*BabelFish*	×	×	×
D–*BabelFish+q*	×	×	√
D–*d*	√	×	×
D–*dq*	√	×	√
D–*do*	√	√	×
D–*doq*	√	√	√
T-*mono*	×	×	×
T-*mono+q*	×	×	√
D-*mono*	×	×	×
D-*mono+q*	×	×	√

Table 5.7: Run descriptions of English–Chinese CLIR experiments.

for use of BabelFish. Examination of the translations gives the explanation. In the NTCIR−3 queries, many of the OOV terms were incorrectly translated syllable by syllable into completely wrong terms. This resulted in a number of incorrect terms being added to a relatively short query. The effect of this was often that many incorrect documents were retrieved. In the NTCIR−4 query set, we observed that in the majority of cases where BabelFish was unable to translate an OOV term, it was simply omitted from the translation. In many cases, there was still enough information in the other query terms to retrieve some relevant documents, especially at high levels of recall.

5.4.2 Disambiguation combined with OOV translation

As shown in Table 5.7, the OOV translation runs T−*do* and D−*do* combine translation disambiguation and OOV detection techniques. The results showed that our OOV translation technique provided an improvement of 18.4% and 15.4% respectively compared to the runs T−*d* and D−*d*

RunID	Average precision	Recall	P@10
T–*BabelFish*	0.1696	0.5107	0.2983
T–*BabelFish+q*	0.1906	0.6113	0.3000
T–*d*	0.1459	0.5239	0.2534
T–*dq*	0.1830	0.5822	0.3000
T–*do*	0.2166	0.5811	0.3448
T–*doq*	0.2386	0.6226	0.3759
T–*mono*	0.2473	0.6209	0.4017
T–*mono+q*	0.2490	0.6458	0.3931
D–*BabelFish*	0.1226	0.4253	0.1828
D–*BabelFish+q*	0.1332	0.4647	0.2155
D–*d*	0.1381	0.5506	0.2569
D–*dq*	0.1678	0.6098	0.3000
D–*do*	0.1932	0.6091	0.3293
D–*doq*	0.2147	0.6468	0.3534
D–*mono*	0.2186	0.6267	0.3603
D–*mono+q*	0.2214	0.6398	0.3397

Table 5.8: Effect of translation disambiguation, OOV term translation, and post-translation query expansion on Chinese–English CLIR; separately and in combination. Average precision, recall, and P@10 for the titles and descriptions of NTCIR−4 Chinese topics. (Data consists of the NTCIR−4 English collection.)

that only applied disambiguation. This improvement was statistically significant at the 95% confidence level, and emphasizes the importance of a good OOV translation technique. The rigid relevance assessment average precision values for title and description runs were 0.2166 (T−*do*) and 0.1932 (D−*do*), respectively representing 87.6% and 88.4% of mono-lingual retrieval effectiveness.

5.4.3 Post-translation query expansion

With the addition of query expansion as described in Section 5.1, results were further improved in all cases. For disambiguation only (no OOV translation) combined with query expansion, our T−*dq* run result (0.1830) was slightly lower than those obtained by applying query expansion to the BabelFish T−*BabelFish+q* results (0.1906), while our D–*dq* run results (0.1678)

were higher. More importantly, combining components of our technique — disambiguation, OOV translation and query expansion, produced results that were statistically significantly higher than those obtained by applying query expansion to BabelFish, for both T−*doq* (0.2386) and D−*doq* (0.2147) runs. The title run (T−*doq*) achieved 95.8% of the mono-lingual query expansion run (T−*mono*+*q*) and description run (D−*doq*) achieved 97.0% of the mono-lingual query expansion run (D−*mono*+*q*).

Although query expansion only gave improvements of 0.8% and 1.3% for the mono-lingual runs, it provided improvements of 10% and 11% for cross-lingual post-translation query expansion runs. This shows that query expansion can usefully improve retrieval effectiveness for imperfectly translated queries, even though it was not helpful for mono-lingual retrieval.

Kwok et al. [2004] showed that using only post-translation query expansion provided the best Chinese–English CLIR results. Similarly, our post-translation query expansion technique provided an improvement of 10% and 11% of title and description runs respectively.

5.4.4 Translation quality

Our CLIR results were close to our mono-lingual benchmark. In comparison to other participants of this task in NTCIR−4, we obtained the second highest results at high levels of precision. The PIRCS retrieval system from City University [Kwok et al., 2004] achieved better overall results. Interestingly their mono-lingual benchmark was higher then ours: 0.3175 and 0.3055 for rigid title and description runs. This suggests the underlying search engine retrieval effectiveness was superior to ours. In CLIR runs they only achieved 75% and 73% of mono-lingual retrieval effectiveness for rigid assessment [Kwok et al., 2004].

This shows that our OOV translation technique has been effective in detecting Chinese OOV terms and extracting English translations, and thus significantly improves CLIR effectiveness. Table 5.3 shows the translations we have extracted from the web. Among 66 potential Chinese OOV terms, 38 instances can be translated word by word using the translation dictionary. Of 28 Chinese OOV terms, we were able to successfully translate 20. The remaining 8 cases failed for one of two reasons: first, our search technique did not return any English terms associated with some Chinese OOV terms; second, some personnel names that relate to events are no longer topical and could not be found on the web, such as "花蝴蝶" (Flojo). However, a system which periodically crawls the web to discover new terms would overcome this problem.

We compared our results to those of the LiveTrans system [Cheng et al., 2004]. The LiveTrans system returns a list of up to 20 alternative translations. In 21 cases the most appropriate

translation were present somewhere in the list. In only 5 cases was the top-ranked translation the most appropriate. Our system only returns the most appropriate translations, and thus **is** correct in 20 instances. In 3 cases our system produced the translations that might be considered more appropriate than LiveTrans. For example, for the Chinese OOV terms "胚胎幹細胞", "基因治療", and "非接觸式智慧卡", our system extracted the English translations "Embryonic Stem Cell", "Genetic Treatment" and "Contactless Smart Cards CSC", respectively; whereas LiveTrans extracted "embryonic stem/stem cells", "gene therapy" and "contactless/smart card" in each case.

5.5 Summary

We experimented with the techniques developed for OOV term translation, translation disambiguation, and post-translation query expansion in Chinese–English CLIR. The results showed that our OOV term translation technique can provide a statistically significant improvement in the retrieval effectiveness, and can be used to improve Chinese segmentation accuracy. In addition, the web proved to be a rich resource of potential translations for topic-specific terms. We also evaluated a two-stage procedure of selecting additional query terms for post-translation query expansion based on term weighting and word association information. This provided a further improvement of up to 11% in Chinese–English CLIR effectiveness. More importantly, we observed that post-translation query expansion can be used to improve effectiveness, especially for imperfectly translated queries. In conclusion, a combination of these techniques provides a significant improvement in CLIR effectiveness, and allows us to achieve up to 97% of monolingual retrieval effectiveness.

Query ID	Chinese query	Detected Chinese OOV terms	Extracted English translations	Given English translations
001	秋門	秋門	(Not found)	Chiutou
002	約翰走路	約翰走路	Johnnie Walker	Johnnie Walker
003	胚胎乾細胞	胚胎乾細胞	Embryonic Stem Cell	Embryonic Stem Cells
004	葛瑞菲絲	葛瑞菲絲	Griffith	Griffith
	喬納	(Not found)	(Not found)	Joyner
	花蝴蝶	(Not found)	(Not found)	Flojo
005	戴奧辛	戴奧辛	Dioxin	Dioxin
006	麥可喬丹	麥可喬丹	Michael Jordan	Michael Jordan
007	巴拿馬運河	巴拿馬運河	Panama Canal	Panama Canal
	卡杜條約	(Not found)	(Not found)	Torrijos-Carter Treaty
008	威而鋼	威而鋼	Viagra	Viagra
009/025	南韓	南韓	South Korea	South Korea
012	黑澤明	黑澤明	Akira Kurosawa	Akira Kurosawa
013	小淵惠三	(Not found)	(Not found)	Keizo Obuchi
014	環境荷爾蒙	環境荷爾蒙	environmental hormone	Environmental Hormone
017	後天免疫缺乏症候群	後天免疫缺乏症候群	AIDS	AIDS
021	電子商務交易	電子商務	Electronic Commerce	Electronic Commercial
022	起亞汽車	起亞汽車	Kia Motors Corp	Kia Motors
030	動物複製技術	複製	clone	Cloning
034	東京都知事	(Not found)	(Not found)	Tokyo provincial governor
038	奈米科技	奈米科技	Nanotechnology	Nanotechnology
046	基因治療	基因治療	Genetic Treatment	Genetic Treatment
048	國際太空站	國際太空站	ISS	International Space Station
051	隱形戰鬥機	隱形戰鬥機	stealth fighter	Stealth Fighter
052	皇太子妃	(Not found)	(Not found)	Crown Princess
	雅子	(Not found)	(Not found)	Masako
053	網際網路	網際網路	Internet	Internet
058	非接觸式智慧卡	非接觸式智慧卡	Contactless Smart Cards CSC	Contactless SMART Card

Figure 5.3: Extracted English translations of Chinese OOV terms, using NTCIR−4 query set.

Chapter 6

Automatic Lexicon Construction for Topical Terms

In English–Chinese and Chinese–English CLIR, the accuracy of translation is limited by the presence of OOV terms. The problem of translation is particularly acute for topical terms, as unfamiliar English terms are typically absent from static lexicons. (In the context of querying, OOV terms are usually compound names or short phrases rather than individual words.) Indeed, translation is also a problem for Chinese readers, which is why it is common practice in Chinese to give both the English and the Chinese form of words whose translation may be unfamiliar. If we can automatically discover the translations of new words from web text, we can improve dictionary coverage and alleviate the OOV problem. By periodically revisiting the web to discover new translations, dictionary coverage can be continually and automatically updated.

In this chapter, we first consider related background work in automatic construction of Chinese–English translation lexicons for OOV terms. Following that we describe the architecture of our system, AutoLex, which extracts topical translations from Chinese text on the web based on the Chinese authorial practice of including unfamiliar English terms in both languages. These are used to construct a Chinese–English lexicon with terms that cannot be found using standard dictionary resources. For some English terms, there are many candidate instances in web documents where the term is given in both languages, but the translations and usage are often inconsistent; in these cases, we show that co-occurrence statistics can be used to identify good translations. In other cases, there are only a few examples of translation for a given term; we show that these too can be used, by making use of the structure of Chinese text. In Section 6.4, we measure the effectiveness

of our system in two ways: human assessment of the quality of translation extracted, and their usefulness in query translation, based on NTCIR and TREC data. We show that it is feasible to crawl the web to build large-scale high-quality bilingual lexicons of OOV terms that do not usually occur in dictionaries, and thus cannot be easily translated in any other way.

6.1 Related Work

Automatic construction of a bilingual translation lexicon has attracted much interest in CLIR research. Chen et al. [2002] proposed a method that uses multiple linguistic resources — including English–Chinese sense-tagged corpora, English–Chinese thesauruses, and a bilingual dictionary — to build a Chinese–English lexicon for translingual applications. Farwell et al. [1992] proposed a method for extracting information from an online resource and using the information to construct lexical entries for a multi-lingual machine translation system. They automatically generated lexical entries for interlingual concepts corresponding to nouns, verbs, adjectives, and adverbs. However, these techniques rely on existing resources, such as parallel corpora or dictionaries, to extract translation pairs; OOV terms are not handled.

Several CLIR researchers have used the web as a resource for OOV translation. McEwan et al. [2002], Yang and Li [2002], and Chen and Nie [2000] attempted to locate parallel documents on the web, and used these to build bilingual dictionaries. However, such approaches suffer from lack of sufficient high-quality parallel texts and limited domain-specific vocabulary. Yang and Li [2002] successfully mined parallel Chinese–English documents from the web, but, as is common with parallel mining, considered only a small domain — press releases from the Hong Kong Government. Cheng et al. [2004] developed the LiveTrans system, which uses anchor text and search results to extract translations of unknown query terms. Each query is submitted to a web search engine to collect the top search result pages, and then co-occurrence and context information between queries and translation candidates is used to estimate their semantic similarity and determine the most likely translations. Wang et al. [2004] proposed a technique for identifying OOV terms within a Chinese corpus prior to query time, and then uses web search engines to determine translations for unknown terms by mining bilingual search-result pages obtained. These approaches can enhance a domain-specific bilingual lexicon. However, they used the web only to extract translations for OOV terms correctly identified in the limited-domain Chinese collections, such as TREC, NTCIR, or medical documents from the STICNET database[1]. This has led some researchers to question how to build a Chinese–English lexicon for topical terms for generalized web searching, in contrast

[1] http://sticnet.stic.gov.tw/

to limited-domain test collections.

Our goal is to build a bilingual translation lexicon of OOV terms that are not domain-specific. Thus, we explore the practicality of using the web as the primary source of OOV terms. In contrast to the method of Wang et al. [2004], we do not need a Chinese corpus, but instead crawl promising web sites.

6.2 The AutoLex Architecture

In this section, we show how the web can be used to construct a large-scale translation lexicon of Chinese–English OOV terms, based on the Chinese authorial practice of including unfamiliar English terms in both languages. Our approach stems from the observation that when new terms, foreign terms, or proper nouns are used in Chinese web text, they are sometimes accompanied by the English translation in the vicinity of the Chinese text.

Our AutoLex system consists of three stages: web site selection, web page processing, and OOV translation extraction. We aim to be as domain independent as possible. Rather than using a specific corpus to identify OOV terms, we mine the web as broadly as possible. We fetched 300 web pages from The Age World[2] (on September 3rd, 2004) and identified a total of 56 English OOV terms as shown in Appendix A. We started by using these 56 randomly selected English seed terms to locate Chinese web sites that seemed likely to contain a mixture of Chinese and English text. Once new OOV terms had been extracted from these sites, we used them to locate further web sites. We then mined these sites to extract bilingual snippets and combine syntactic structure analysis with co-occurrence statistics to infer translations. An added advantage of this approach is that, by first detecting English OOV terms and then extracting Chinese translations, we avoid problems associated with Chinese segmentation — a particular problem in the presence of OOV terms, since there is no prior segmentation information available.

6.2.1 Web site selection

We used several web search engines[3] to fetch up to 100 top-ranked Chinese web pages each, using the 56 randomly selected English OOV terms as queries. We expected that there would be considerable overlap among the sites returned by the different search engines. The overlap is depicted graphically in Figure 6.1, and interestingly is not as high as might be anticipated; for example, of the 3054 pages found by Baidu, only 906 are found by the other engines. We note

[2] `http://www.theage.com.au/news/world/`

[3] `www.google.com`, `www.yahoo.com.cn`, and `www.baidu.com`.

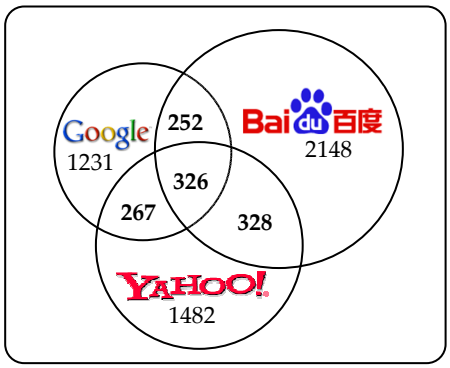


Figure 6.1: Overlap among the sites returned by the different search engines — Google, Baidu, and Yahoo. The number in each region is the number of returned web sites in that category.

that some researchers have only used top documents from only one or two search engines when determining translations at run time [Chen et al., 2000; Cheng et al., 2004]. From our results we conclude that several search engines should be used.

Using these queries and search engines, 13,194 web pages from around 6034 distinct web sites were collected. The web sites were then ranked according to the number of distinct web pages retrieved. Our hypothesis was that the higher-ranked sites would be more likely to provide good numbers of OOV terms. To test this, we randomly selected 100 English OOV terms from the 89,848 we extracted from the 80 GB of web text and used them to collect and re-rank the web sites. The new ranking produced 11,507 distinct sites, compared to the 6034 sites originally collected. There were 44 sites that appeared in the top 100 ranked list from the previous run. From this we can see that the web text we have collected represents only a small portion of relevant Chinese sites. However, the degree of overlap in the top 100 ranked sites suggests that our selection technique is reasonably robust.

Having access to only limited resources, we were unable to crawl all of the sites. We therefore decided to crawl as many of the highly ranked sites as we could, in order to maximize the number of OOV terms that we might extract. We also crawled some lower-ranked sites. The purpose of this pilot study is to explore the correlation between site rank and number of OOV terms extracted. (However, we observed that there is only a weak correlation between these two factors.) In total, we collected 80 GB of web text.

6.2.2 Web text processing

The web data was filtered by an HTML parser and segmented using the punctuation delimiters to collect lines containing English strings. The Chinese–English snippets we have collected from this process fell into three kinds:

Case 1. Chinese text was followed by the English strings without any quotation marks. For example:

... 俄罗斯的能源巨头 Gazprom 公司 ...
... 名誉侵权案 CNNIC 告赢 3721 ...

In this case, the English strings "Gazprom" and "CNNIC" are used directly in Chinese web pages. We found that less than 1% of English strings that occurred without brackets were accompanied by their Chinese translation, and thus we did not use this pattern as a source of OOV translations.

Case 2. Chinese text within various types of Chinese quotation marks was followed by English strings within brackets. For example:

... 最新"迅驰"(Centrino)移动技术 ...
... 把华人社区称为'华埠'(China Town) ...
... 发现的"成纤维细胞生长因子（FGF）" ...
... 网络游戏《仙境传说(Ragnarok)》的战斗 ...

We extracted 27,590 such instances from 80 GB of web text; our evaluation (reported later) showed that 98% of these were correct translation pairs. The high accuracy is both because this format is a convention commonly used to indicate translations and also because the use of quotation marks eliminates problems associated with Chinese word segmentation.

Case 3. Some Chinese text was followed by English strings in brackets, for example "嫌疑人声称自己是摩洛伊斯兰解放阵线 (MILF) 成员". Superficially, this pattern looks similar to the one discussed above. However the lack of delimiters in the Chinese text means that more sophisticated translation extraction techniques must be used.

6.2.3 OOV translation extraction

The OOV translation extraction algorithms for case-3 terms employ co-occurrence statistics using the frequency and length of each Chinese substring. We have observed that there are only a small number of syntactic structures that occur frequently in Chinese text that can be used to infer correct translation, especially when no other co-occurrence evidence is available. Essentially, these words are also acting as delimiters, demarcating the start of the Chinese string that is the correct starting point of the Chinese translation. For example, in:

... 那名队员[名叫]查尔斯戈雷弗 (Charles Graver) ...
... [比如说]西门子 (Siemens) ...
... 又被[称为]数据库知识发现 (KDD) ...
... 大名鼎鼎[的]占克顿伯里圈 (Chanctonbury Ring) ...

the Chinese words "名叫"(is called), "比如说" (for example), and "的" ('s) are syntactic markers. By correctly identifying these markers we were able to discard all Chinese characters prior to and including the marker. Thus the strings in the above example can be shortened to:

查尔斯戈雷弗 (Charles Graver)
西门子 (Siemens)
数据库知识发现 (KDD)
占克顿伯里圈 (Chanctonbury Ring)

Such truncation greatly simplifies the issue of determining the correct word boundary of Chinese translations. We collected a total of 208,379 Chinese–English case-3 occurrences from 80 GB of web text, and 47% of them contained these syntactic structures. This technique improves translation extraction effectiveness even when the co-occurrence frequency is low.

After using the above technique, for each distinct case-3 English string we collected the frequency and the length of all immediately preceding Chinese substrings. For example, given the string "占克顿伯里圈 (Chanctonbury Ring)", we collect the Chinese substrings "里圈", "伯里圈", "顿伯里圈","克顿伯里圈", and "占克顿伯里圈" as candidate translations of the English string "Chanctonbury Ring". The Chinese substring with the highest weight was then selected as the most appropriate translation of the English string. However, as we now discuss, weighting is not straightforward. In general, frequent substrings are preferred over those that occur less frequently, and longer substrings are preferred over shorter. However the natural distribution is that shorter

strings occur more frequently than longer ones. In most cases, a weight that is the product of the length and the frequency of each substring will suffice.

Sometimes, however, an English term co-occurs with several Chinese strings of different length. The simple length-frequency product gives a similar weight in both cases, although in some situations the short string is the correct translation whereas in other cases the long string is the correct translation. Inspection of the data showed two distinct patterns. In one situation, the correct translation was a short substring that was common to all these strings. The rest of the characters where related, but were not part of the translation. For example in "日本游戏软件公司世嘉 (Sega)", "世嘉" is the Chinese translation of the English term "Sega" where "日本游戏软件公司" (Japanese game software company) is noise.

In the other situation, typically the transliteration of a foreign name such as "伊恩索普" (Ian Thorpe), sometimes only the surname "索普" co-occurs with the English string, and sometimes the full name "伊恩索普" is given. In this situation the long string is the correct translation. We also noticed that, in the first situation, the short string typically occurs with a high frequency, and by placing more importance on frequency in this situation we can usually select the correct translation. Conversely, by placing more importance on length in the second situation, we are able to select the long string as the correct translation. Thus we used one method of calculating the weight when $f_t/f_e > \alpha$, and otherwise used another method, where α is a parameter:

$$w_t = \begin{cases} f_t \times l_t + f_t & \text{if } f_t/f_e > \alpha \\ f_t \times l_t + l_t & \text{otherwise} \end{cases} \tag{6.1}$$

Here l_t is the length of the Chinese substring, f_t is the co-occurrence frequency of the Chinese substring and the English string, and f_e is the frequency of the English string.

While wishing to avoid over-tuning the formula, we nonetheless tested several values of α using case-3 data, in order to examine the sensitivity of the algorithm to this parameter. The results are shown in Table 6.1. It can be seen that a value of $\alpha = 0.65$ gives slightly better results, compared to $\alpha = 0.5$ and $\alpha = 0.8$. Lower values of α lead to an increase in the number of incomplete translations selected, while higher values lead to more correct translation with additional misleading terms. As explained above, the effect of adding either f_t or l_t in this way is that the greater the difference between f_t and l_t, the greater the adjustment to the weight. This has the effect of biasing the selection towards strings that occur with high frequency or are long. In order to show the effect of this bias, we have also included a column in Table 6.1 that shows the translations extracted if only $f_t \times l_t$ is used as the weighting function. As expected, the results are inferior.

In order to investigate the effect of using syntactic structure on the translation correctness for

	Baseline using $w_t = f_t \times l_t$	if $f_t/f_e > \alpha, w_t = f_t \times l_t + f_t$ otherwise, $w_t = f_t \times l_t + l_t$		
		$\alpha = 0.5$	$\alpha = 0.65$	$\alpha = 0.8$
1. **Exactly correct translations**	60	62	63	63
Partically correct translations				
2. with 1 or 2 related extra words	11	14	16	14
3. with more than 2 related extra words	5	6	4	6
4. with any other misleading words	11	6	5	8
5. but incomplete	8	8	6	4
6. **Wrong translations**	6	6	6	5
1 + 2 + 3	**76**	**82**	**83**	**83**

Table 6.1: Effect of α on translation extraction effectiveness for case-3 terms. (All values are percentages.)

low-occurrence frequency translation pairs, we experimented with using a set of different threshold values for the co-occurrence frequency of case-3 data. As shown in Table 6.2, the translation correctness for low-occurrence frequency translation pairs was at least as good as for those pairs occurring with high frequency. Table 6.2 also shows the important contribution of syntactic structure analysis, improving the proportion of correctly extracted translations from 33% to 63%. The use of syntactical structure has allowed us to greatly increase the accuracy of translation in circumstances where the co-occurrence frequency of translation pairs is low, and thus greatly increase the size of automatically-created dictionary.

6.3 Effectiveness of Web Crawling

As discussed in Section 6.2.1, one starting assumption was that domains that returned more distinct pages in response to our initial queries would on average provide more OOV translation pairs. Figure 6.2 shows this relationship, and it can be seen that there is only a weak correlation between these two factors. We also investigated the relationship between the size of the web sites we crawled and the number of OOV terms extracted. This data is shown in Figure 6.3. Again, it can be seen that there is a weak correlation. Obviously, as more data is collected, the number of OOV terms extracted increases. We expected that as more data is collected the rate of appearance

	with syntactic structure analysis				
	No &	*Yes &*			
	$f_t \geq 1$	$f_t \geq 1$	$f_t \geq 3$	$f_t \geq 5$	$f_t \geq 7$
1. **Exactly correct translations**	33	63	64	66	64
Partially correct translation					
2. with 1 or 2 related extra words	19	16	11	9	5
3. with more than 2 related extra words	9	4	1	1	2
4. with any other misleading words	26	5	2	2	1
5. but incomplete	4	6	17	17	22
6. **Wrong translations**	7	6	6	6	6
Total number of OOV terms extracted	71,324	69,171	9748	4062	2482

Table 6.2: Effect of using syntactic structure on translation extraction effectiveness for case-3 terms. (All values are percentages.)

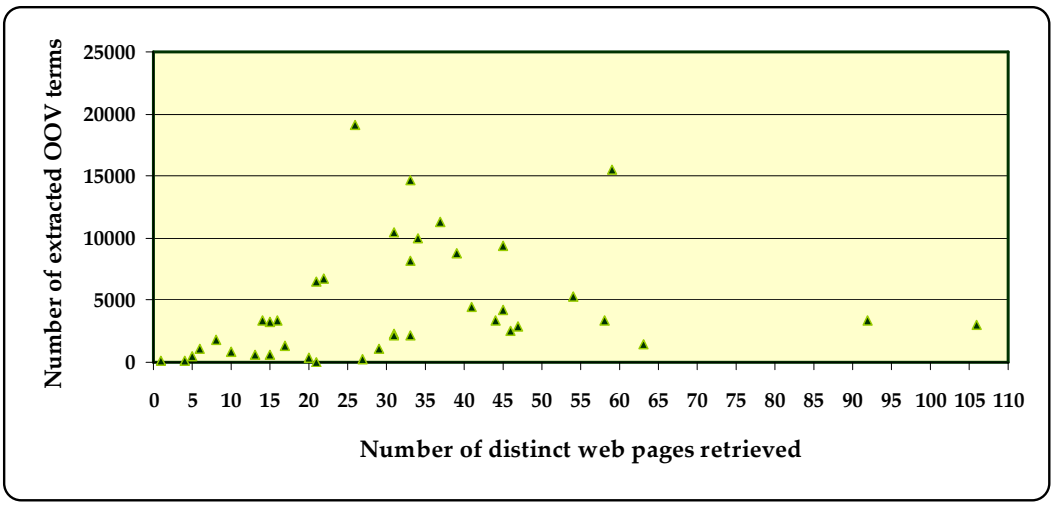


Figure 6.2: Web site rank versus the number of extracted OOV terms.

of new terms would decrease, and possibly start to asymptote. To investigate this relationship, six different sized web collections were constructed. Starting from a 1.25 GB web collection, the size was successively doubled up to a maximum of 80 GB, from which we extracted a total of 89,848 new translation pairs. This is shown in Figure 6.4. We have nearly doubled the baseline dictionary, an LDC English–Chinese bilingual wordlist containing 110,834 entries.

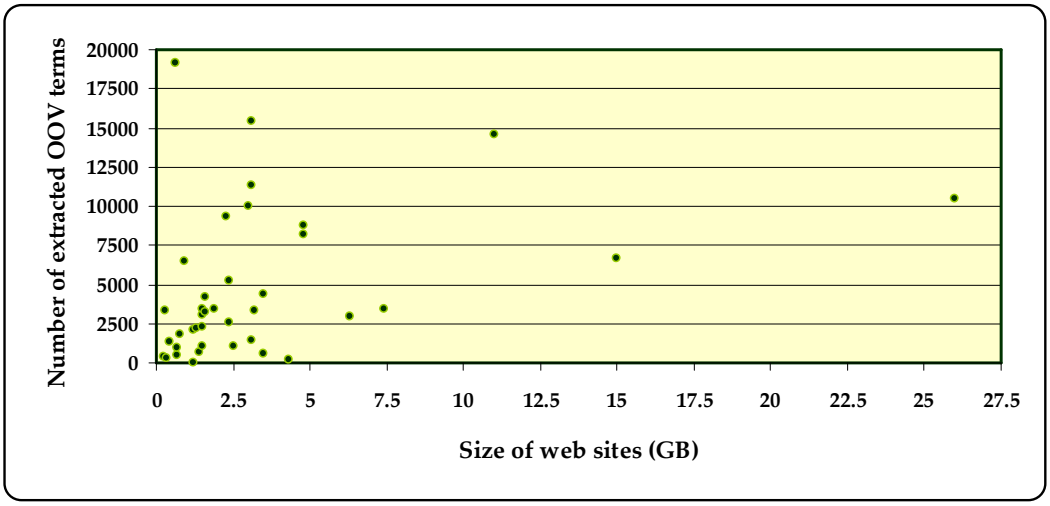


Figure 6.3: Web site size versus the number of extracted OOV terms.

It can be seen that the rate of appearance of new OOV terms is indeed decreasing but has not yet started to asymptote. Estimates of the size of the web, and of the Chinese web, vary. In 2005, Google claimed that they have indexed over 8×10^9 web pages. Estimates of the number of Chinese pages dating from 2002[4] and 2003 [Gey et al., 2005] are 48×10^6 and 93×10^6 respectively, and so we assume that the figure may currently be in the region of 200×10^6. Our crawl of 80 GB of data contains 400,000 pages and thus represents only about 0.2% of the likely total of the Chinese web. Therefore our sample is too small to allow us to reliably estimate the total number of OOV terms that can be extracted from the Chinese web.

Figure 6.5 shows the relationship between volume of data and the number of translations for which new meanings were extracted, as well as the number of existing meanings extracted (where a translation was given for a term in the dictionary we use). It can be seen from this graph that the vast majority of in-vocabulary terms extracted hold new meanings. In Section 3.3.4, we showed that one of the reasons for failure of CLIR query translation is lack of appropriate alternative translations in bilingual dictionaries.

6.3.1 Frequency of crawling

We hypothesis that it should be possible to update an existing English–Chinese translation lexicon by periodically crawling the web. For example, once-novel terms such as "Clinton" are no longer explicitly translated for Chinese readers; without periodic update such translations will be lost.

[4] `http://www.netz-tipp.de/languages.html`

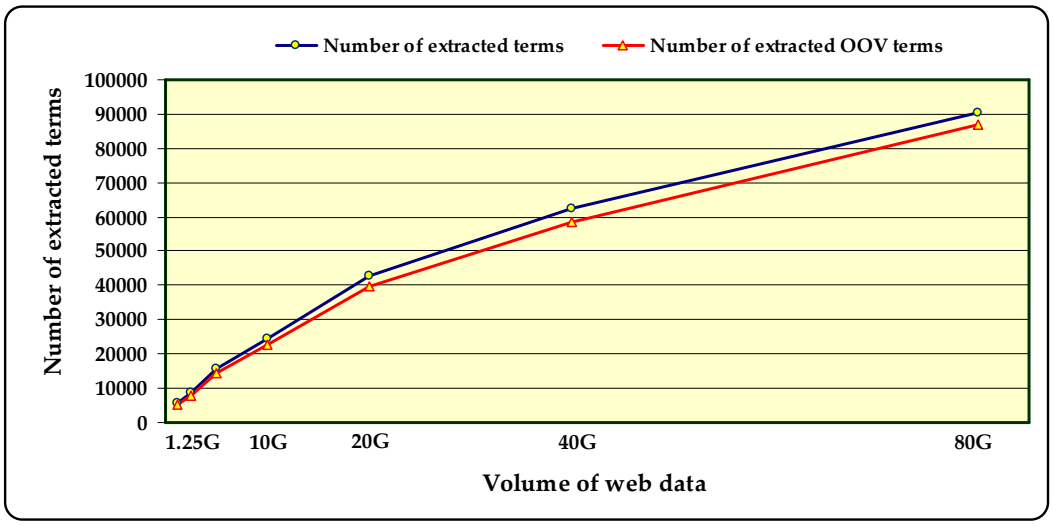


Figure 6.4: Volume of web versus the number of extracted out-of-vocabulary terms.

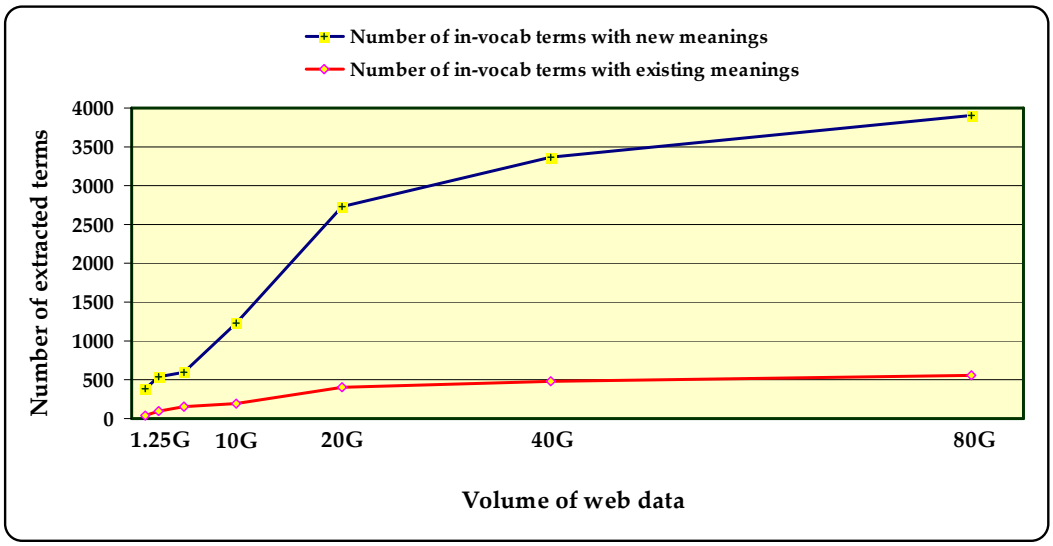


Figure 6.5: Volume of web text versus the number of extracted in-vocabulary terms.

We did not extensively investigate this issue, but did examine it. We re-crawled two news web sites (`news.sina.com.cn` and `news.sohu.com`) after an interval of two months to measure the change in the number of translations pairs extracted, as shown in Table 6.3. The *processed* column refers to the amount of data remaining after HTML markup and other non-content material were removed. For the web sites selected, it can be seen that there was a significant addition to the

	Data size		Number of extracted terms				Timestamp
	Raw (GB)	Processed (MB)	OOV	In-vocabulary terms with new meanings	In-vocabulary terms with existing meanings	Total	*from — to*
`news.sina.com.cn`							
1st-crawl	3.1	118	1445	65	12	1522	*12/2001–09/2004*
2nd-crawl	4.3	127	2325	107	30	2462	*07/2002–11/2004*
`news.sohu.com`							
1st-crawl	7.4	191	3390	148	36	3574	*06/2001–09/2004*
2nd-crawl	12.0	223	4116	189	44	4349	*06/2001–11/2004*

Table 6.3: Effect of revisiting web sites on OOV term discovery.

number of OOV terms extracted as well as new meanings for existing terms. This shows that construction of a bilingual OOV lexicon requires periodic updating to capture new terms as they enter the vocabulary. A full implementation of this extraction technique would need to consider how frequently such major sites should be re-crawled. However we noted that many news-based sites file articles by date. An intelligent crawling system that could learn the structure of different sites would only need to crawl newly added material.

6.3.2 Re-ranking of web sites

At the start of our investigation we used 56 randomly selected OOV terms to seed our process and collect a set of web sites to rank and subsequently mine, as discussed in Section 6.2.1. We investigated whether our initial queries biased our process towards a particular portion of the web to the exclusion of other parts, or whether the web site selection technique is able to find the majority of useful pages.

To test this, we randomly selected 100 OOV terms from the 89,848 we extracted from the 80 GB of web text and used them to collect and rank web sites as we had done before. The new ranking produced 11,507 distinct sites, compared to the 6,034 sites originally collected. There were 44 sites that appeared in the top 100 ranked list from each run. From this we can see that the web text we have collected represents only a small portion of relevant Chinese sites. However, the degree of overlap in the top 100 ranked sites suggests that our selection technique is reasonably robust.

6.4 Effectiveness of Translation

Our aim in this study is to develop robust techniques to construct a translation lexicon for otherwise OOV terms. Although such a lexicon may ultimately be used for CLIR, we do not evaluate our work in terms of CLIR effectiveness. Previous studies that used test collections as a starting point can identify every OOV term in the set of test queries and explicitly search from these on the web. If a translation exists for the known OOV term, the major issue is whether it can be identified and correctly extracted.

However, in a production system one cannot anticipate the queries that users will enter. That is why, in this research, we have taken the approach of combing the web to build up a translation lexicon. However, due to resource limitations we have only covered a small portion of the web, and hence only found a portion of the available OOV translations. Thus any evaluation of CLIR effectiveness using queries from a standard test set would inevitably produce inconclusive results, as a consequence of the limited coverage rather than anything to do with the accuracy of the technique. We therefore use two methods of evaluation: human assessment of translations, to measure their correctness, and lexicon usefulness.

6.4.1 Accuracy of extracted translations

We used a native Chinese speaker (not associated with this research) to evaluate the translations we extracted from the web. From 89,848 extracted OOV translation pairs, 400 pairs were randomly selected, giving an error margin of ±5%. The person was given the web pages, a web search engine for further validation, and the extracted English–Chinese translation pairs. Results are shown in Table 6.4.

We can see that 63% of the translations found were strictly correct. However, an additional 17% of translations contained the correct translation, with some additional related information, such as the person's title or the organization's location. When used for retrieval, this additional information may arguably improve effectiveness. A further 7% of translations were correct but included some misleading terms, or terms commonly associated with the OOV term. For example "第二届美国偶像大赛" is extracted as the translation of "American Idol", where term "the second" (第二届) is commonly associated with the OOV term "American Idol" (美国偶像大赛) in web documents. In 7% of cases involving translation of personal names, only the surname was found, for example "克莱维茨" is extracted as the translation of "Lenny Kravitz". Arguably in both of these cases such translations would find relevant documents although the precision may be lower. Only 6% of translations were assessed as wrong. Additionally, the proportion of correct

	Proportion
1. Exactly correct translation	63
2. Correct translation with 1 or 2 related extra words	16
3. Correct translation with more than 2 related extra words	1
4. Correct translation with any other misleading words	7
5. Correct translation but incomplete	7
6. Wrong translation	6

Table 6.4: Accuracy of extracted translations. (All values are percentages.)

Kind of name	**Proportion**
Personal names	28.65
Organizational names	14.06
Technical terminology	17.45
General phrase	11.72
Movie titles	7.29
Abbreviation	6.25
General term with new meanings	3.65
Place name	2.60
Military terms	2.08
Brand names	2.08
Song titles	1.82
Book titles	1.30
Music band	0.52
Games	0.52

Table 6.5: Distribution of the OOV translations extracted from 80 *GB of web text. (All values are percentages.)*

translations (with one or two related extra words) improved significantly, from 5% to 16%.

We also undertook a simple analysis of the types of OOV terms that we extracted. This breakdown is shown in Table 6.5. It can be seen that personal names, names of organizations, and technical terminology comprise 60% of the OOV terms. There are also phrases, movie titles, and abbreviations. To a large extent, these are not terms that we would expect to find in a static dictionary.

We compared our results to those of the LiveTrans system [Cheng et al., 2004]. However although this system is the most similar to ours, the two systems are not easily comparable. First, LiveTrans searches the web at run time, an activity which we aim to eliminate. Second, the LiveTrans system returns a ranked list of up to 20 alternative translations, whereas Autolex returns only the most likely one. To test the accuracy of the extracted translations returned by Livetrans, we used a total of 57 OOV terms observed in the NTCIR (3 and 4) and TREC (5 and 6) queries. Of these, 14 exactly correct translations are top-ranked translations extracted by LiveTrans. When using OOV terms from news topic queries, LiveTrans has an accuracy of around 25%, whereas, for an unrestricted domain, greater than 60% of the AutoLex translations are exactly correct. If one or two related words are tolerated, correctness is almost 80%.

6.4.2 Lexicon usefulness

Estimates of the proportion of queries that contain OOV terms vary significantly, and this obviously depends on the particular corpus and query set. To evaluate the usefulness of our technique, we observed the proportion of OOV terms in various query sets that can be translated using the OOV translation lexicon we have constructed.

One measure we used was to see how many OOV terms from various NTCIR (3 and 4) and TREC (5 and 6) query sets we were able to find. In the NTCIR queries, we previously observed a total of 32 OOV terms. AutoLex was able to locate 16 of these terms and their exactly correct translations, and LiveTrans was able to correctly translate 20 terms, of which 9 were top ranked. There were 25 OOV terms observed in the TREC queries. LiveTrans correctly translated 20 English OOV terms, of which 5 were top ranked, and AutoLex located a total of 5 OOV terms.

Although AutoLex was not as effective as LiveTrans (run-time OOV translation extraction), this is not surprising given our limited coverage of the web. First, the goal in our work is to extract OOV translation pairs independently of any specific collections, allowing us to carry out the process prior to query time, and thus improve the efficiency of query translation. Second, although we have collected 80 GB of Chinese web text, this represents only around 0.03% of the Chinese web and a limited time frame. Greater coverage will increase the proportion of OOV terms found. Indeed, it is surprising that we found so many of the NTCIR and TREC OOV terms in such a small portion of the Chinese web, especially given that these terms are no longer particularly current. That is, the test queries from TREC and NTCIR relate to events in the mid- to late nineties, whereas most web sites only contain translations from more recent events.

6.5 Summary

In this chapter, we present an automatic system, AutoLex, which uses the web to build a Chinese–English translation lexicon of topical terms, based on the Chinese authorial practice of including unfamiliar English terms in both languages. Inconsistencies in how the translations are presented lead to ambiguity; however, we showed that these can be resolved statistically. Of the Chinese–English text snippets we collected, the great majority were cases where the Chinese text was not punctuated, and thus there was no indication of where the translation began and finished; and most of these only occur once. The use of syntactical structure has allowed us to greatly increase the accuracy of translation in circumstances where the co-occurrence frequency of translation pairs is low, and thus greatly increase the size of automatically-created dictionary.

The extracted translations are shown to be reasonably accurate. Such translations may be added to both English-to-Chinese and Chinese-to-English dictionaries, that can latter be used at query time, thus improving the efficiency of query processing. Moreover, such a system has wider applications than CLIR; for example, discovered translations of OOV terms can guide development of printed translation dictionaries and can be used in machine translation of full text.

We have also shown that our techniques for locating web sites likely to contain a high proportion of OOV terms work reasonably well. If external search engines are used as part of the site locating process in query-time translation, it is essential that several different engines be used to rank the most useful sites, as we have shown that the overlap for the top 100 pages retrieved is low.

Chapter 7

Automatic Acquisition of Web Parallel Corpora

A parallel corpus is a collection of texts that have been manually translated into two or more languages by human translators. Parallel corpora provide a rich source of translation information. They have been used to train statistical translation models [Nie et al., 1999; Franz et al., 2001; Brown et al., 1990], translation disambiguation systems [Ballesteros and Croft, 1998], OOV term translation [McEwan et al., 2002], and multilingual thesaurus construction [Chau and Yeh, 2001]. However, some parallel corpora are subject to subscription or licence fee and thus not freely available, while others are domain specific. For example, parallel corpora provided by the Evaluations and Language resources Distribution Agency[1], the Linguistic Data Consortium[2], and the University Centre for Computer Corpus Research on Language[3], all require subscription or fee. There are several large manually constructed parallel corpora available on the web but they are domain specific, significantly limiting their practical use in CLIR tasks. Examples include the Bible in a number of languages (collected by the University of Maryland) and the European parliament proceedings parallel corpus (1996-2003)[4] in eleven European languages. In order to take advantage of publicly available parallel corpora, a robust system is needed to automatically mine them from the web.

In this chapter, we present a system to automatically collect high quality parallel Chinese–

[1] http://www.elda.org/

[2] http://www.ldc.upenn.edu/

[3] http://www.comp.lancs.ac.uk/computing/research/ucrel/

[4] http://people.csail.mit.edu/koehn/publications/europarl/

English corpora from the web — Web Parallel Data Extraction (WPDE). In Section 7.1, we consider other related work. Section 7.2 lays out the WPDE architecture. Section 7.3 describes our evaluation methodology. In Section 7.4 we detail our experiments and present the results obtained.

7.1 Existing Parallel Text Mining Systems

The information available on the web presents a valuable new source of parallel text. Recently, several systems have been developed to exploit this opportunity.

Nie et al. [1999] developed the PTMiner to mine large parallel corpora from the web. PTMiner used search engines to pinpoint the candidate sites that are likely to contain parallel pages, and then used the URLs collected as seeds to further crawl each web site for more URLs. The pairs of web pages were extracted on the basis of manually defined URL pattern-matching, and further filtered according to several criteria, such as file length, HTML structure, and language character set. Several hundred selected pairs were evaluated manually. Their results were quite promising; from a corpus of 250 MB of English–Chinese text, statistical evaluation showed that of the pairs identified 90% were correct.

STRAND [Resnik and Smith, 2003] is another web parallel text mining system. Its goal is to identify pairs of web pages that are mutual translations. Resnik and Smith used the *AltaVista* search engine to search for multilingual websites and generated candidate pairs based on manually created substitution rules. The heart of STRAND is a structural filtering process that relies on analysis of the underlying HTML to determine a set of pair-specific structural values, and then uses those values to filter the candidate pairs. Approximately 400 pairs were evaluated by human annotators. STRAND produced fewer than 3500 English–Chinese pairs with a precision of 98% and a recall of 61%.

The Parallel Text Identification System (PTI) [Chen et al., 2004] was developed to facilitate the construction of parallel corpora by aligning pairs of parallel documents from a multilingual document collection. The system crawls the web to fetch (potentially parallel) candidate multilingual web documents using a web spider. To determine the parallelism between potential document pairs, a filename comparison module is used to check filename resemblance, and a content analysis module is used to measure the semantic similarity. The results showed that the PTI system achieves a precision rate of 93% and a recall rate of 96%. PTI is correct in 180 instances among a total of 193 pairs extracted. Our later evaluation showed that WPDE is able to produce 373 correct pairs with a precision of 97% and a recall of 94% on the same domain, using the file length

	Precision	Recall	Parallel text size	Number of pairs evaluated
PTMiner	90%	–	250 MB	100–200 pairs (randomly picked)
STRAND	98%	61%	3500 pairs	400 pairs (randomly picked)
PTI	93%	96%	427 pairs	427 pairs

Table 7.1: Summarized results from PTMiner, STRAND, and PTI

feature verification (as described in Section 7.2.3) only.

A summary of the results from the above studies is tabulated in Table 7.1.

7.2 The WPDE Architecture

WPDE is an automatic system for large scale mining of parallel text from existing English–Chinese bilingual web pages in a variety of domains. In summary, our procedure consists of three steps: selection and crawling of candidate sites, extraction of candidate pairs, and verification of parallel pairs.

7.2.1 Selection and crawling of candidate sites

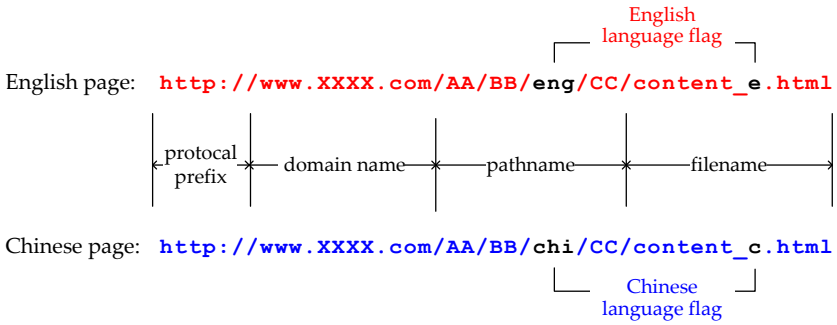
Rather than using search engines to identify the candidate sites, we started with a snapshot of two million web pages from Microsoft Research. We noticed that images representing the language types are almost always accompanied by their text equivalents — ALT text. One of the major differences between WPDE and previous systems is that the candidate sites are selected on the basis of both anchor text and image ALT text. For a given web page, we extract the hypertext links when the anchor text or the image ALT text matches a list of pre-defined strings that indicate English, simplified Chinese, and traditional Chinese (see Appendix B). If a website contains two or more hypertext links to the different versions, we select these as candidate websites. We then selected 1598 candidate websites based on the anchor text, and 211 extra candidate websites were obtained using the image ALT text.

Once candidate sites were extracted from the snapshot, we used `wget`[5] to fetch all documents from each site on the live web and create local copies of remote directory hierarchies.

[5] `http://www.gnu.org/software/wget/`

7.2.2 Extraction of candidate pairs

We then extract candidate parallel pairs from the crawled web pages. URLs consist of a protocol prefix, a domain name, a pathname, and a filename. Webmasters tend to give the pages similar names if they are translations of each other. The only difference between these two URLs is the segments that indicate the language type. For example, given the URLs of an English–Chinese parallel pair,



where `eng` and `e` are used to indicate the English version and `chi` and `c` are used to indicate the Chinese version. We observed that there are only five patterns `e`, `en`, `eng`, `engl`, and `english` that are used to indicate the English version. In contrast, the patterns employed to indicate the Chinese version are quite unpredictable, and it is unrealistic to expect a "complete" pattern list. Therefore, previously employed language flag matching approaches [Nie et al., 1999; Kraaij et al., 2003], which replace one language prefix/suffix/infix with all possible prefixes/suffixes/infixes in the other language based on a static pre-defined pattern list, will not work on a large scale URL matching process.

An improved approach combining pattern matching and edit-distance similarity measure [Wagner and Lowrance, 1975] has been exploited in our work. For example, if an English pattern is detected in the pathname of an URL, we first extract the candidate Chinese URLs with the same protocol prefix, the same domain name, and the same pathname, except for the language flag segment. If the Chinese URL contains a language pathname segment that is in our standard Chinese pattern list (`c`, `ch`, `chi`, `Chinese`), we select this URL. Otherwise we use an edit distance metric to find the nearest match to one of these Chinese patterns, such as `tc`, `sc`, `tchi`, or `schi`. If the filenames are the same, the process is finished. Sometimes an additional filename matching step is required. In the simplest case the filename will differ by one of the standard language flag patterns, otherwise we again use the same edit distance function to find the filename closest to one of these Chinese patterns.

We have extracted a total of 7894 candidate pairs. Later evaluation showed that, in isolation, this approach has a precision of 79%. Among a total of 606 pages, which are in `.pdf`, `.doc`, `.rtf`, and `.cfm` format, 558 of them are parallel pages with a high quality. We would suggest the web documents in these specific formats as a reliable parallel text source.

7.2.3 Verification of parallel pairs

The candidate pairs extracted in the previous steps can be further filtered based on three common features of parallel pages: the file length, the file structure, and the translation of the web page content. To filter out the pairs that are not similar enough, a threshold is set to each feature score. The experimental results are shown in Section 7.4.

File length

We assume the files sizes of Chinese–English parallel texts are roughly proportional. Additionally, files of length 40 bytes or less are discarded. Using these metrics, 323 candidate pairs (5%) were filtered out. For the candidate pairs that remain, we then calculate the ratio of the two file lengths $S_{len} = \text{length}(f_{ch}) \,/\, \text{length}(f_{en})$. This ratio is then used in combination with other features as described below.

File structure

The HTML structures of two parallel pages should be similar. We extract the linear sequences of HTML tags from each candidate pair, then apply case-folding and remove noise, such as `meta`, `font` and `scripts`. Unix `sdiff` is used to find differences between these two sequences of HTML tags. For example, as shown in Figure 7.1, consider the two sequences of HTML tags on the left, the aligned sequence generated by `sdiff` is shown on the right.

The feature score of the file structure is calculated using $S_{struct} = N_{diff} \,/\, N_{all}$, where $N_{diff} = 4$ is the number of unaligned lines in the given example above, and $N_{all} = 12$ is the total number of the lines, and is used to normalize the score. Thus, the lower the score the better, with 0 being ideal.

Content translation

To consider the content translation of a candidate parallel pair, we align the two pages using the Champollion Tool Kit[6], which provides ready-to-use parallel text sentence alignment tools.

[6] `http://champollion.sourceforge.net/`

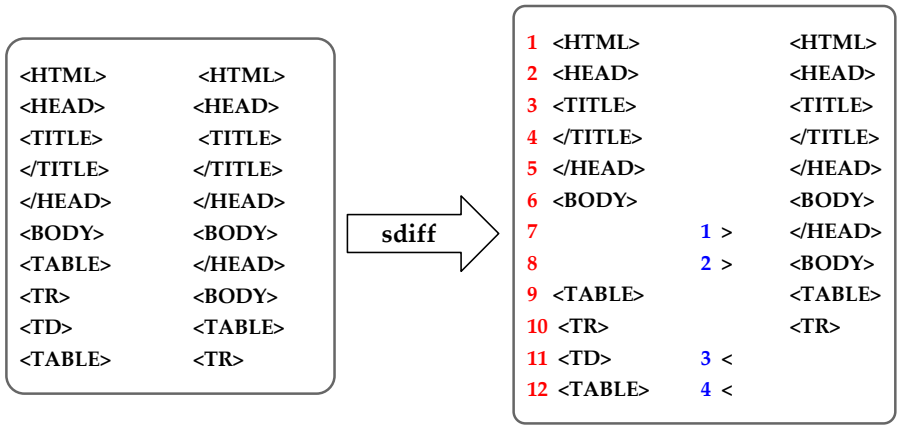


Figure 7.1: An example of file structure comparison using sdiff.

Champollion depends heavily on lexical information, but uses sentence length information as well. Past experiments indicate that Champollion's performance improves as the translation lexicon becomes larger. We therefore compiled a large English–Chinese lexicon, which contains 250,000 entries. The score of the content translation feature is calculated using $S_{trans} = N_{aligned} / N_{(ch,en)}$, where $N_{aligned}$ is the number of aligned sentences and $N_{(ch,en)}$ is the total number of sentences in the two pages.

7.2.4 K-nearest-neighbors classifier

After investigating the recall-precision results of each single feature verification, we observed that, although the file length feature produced the highest precision, the file structure feature can achieve a relatively high recall when lower precision is acceptable. Intuitively, it is possible to achieve better overall performance if multiple features can be combined using an appropriate model. To observe the data distribution in a 2-dimensional feature space, we generated the scatter plot matrix shown in Figure 7.2. The file length feature score is plotted in the X axis, while the file structure feature score is plotted on the Y axis. The 'true' pair is marked by a triangle and the 'false' pair is represented by a cross. As we can see, in the case of a mixture of tightly clustered 'true and false' data, a linear decision boundary is unlikely to be optimal, so that k-nearest-neighbors method would be more appropriate.

KNN has been successfully used for pattern classification on many applications [Cover and Hart, 1967]. Being a non-parametric classification method, it is a simple but effective method for classification. It labels an unknown sample with the label of the majority of the k nearest

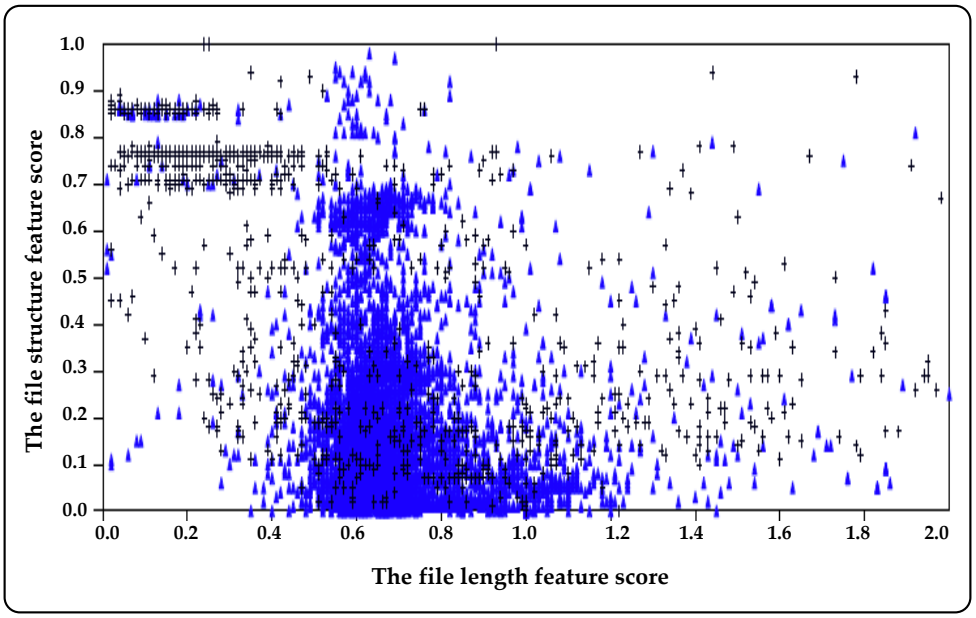


Figure 7.2: A scatter plot of the 2-feature dimensions. The x-axis shows the file length feature score. The y-axis shows the file structure feature score.

neighbors. A neighbor is deemed nearest if it has the smallest distance. The distance is usually calculated using the Euclidean distance.

Using a total of 6500 English-Chinese candidate pairs, we carried out tenfold cross-validation experiments using a KNN classifier to predict the correctness of a candidate pair. Specifically, the data is randomly split into 10 disjoint validation subsets, each with 650 pairs. In each fold, we then select one of those subsets as a test set with 650 test items and use the remaining 5850 pairs as its training set; the fraction of true and false pairs in each fold's test and training sets approximates the overall division, 80% to 20%, respectively. The choice of k affects the performance of a KNN classifier. Wilson and Martinez [1997] proposed that the k is typically a small odd integer and often determined from cross-validation. Therefore we choose the optimal k value with the best performance in cross-validation experiments. Through our experiments, we determined that the best results are generally obtained with $k = 15$ for 3-feature dimensions, and $k = 7$ for 2-feature dimensions.

7.3 Evaluation Methodology

The performance of a system that finds web parallel pages can be evaluated using standard IR measures of precision and recall. Precision represents the proportion of candidate parallel pages retrieved that are correct, thus:

$$Precision = \frac{Number\ of\ correctly\ aligned\ pairs}{Total\ number\ of\ aligned\ pairs}$$

Whereas recall represents the proportion of parallel pages that the system actually found:

$$Recall = \frac{Number\ of\ correctly\ aligned\ pairs}{Total\ number\ of\ parallel\ pairs\ in\ the\ collection}$$

Recall can be calculated for a test collection since the total number of parallel pairs can be determined by inspection, but cannot be calculated for the entire web.

We used three Chinese–English bilingual speakers to evaluate the correctness of all the parallel pairs we extracted from the web. If the English and Chinese pages contain entirely the same meaning, the pair is annotated as a 'correct pair'. While previous systems have been evaluated on relatively small data sets (a few hundreds of pairs), we created a large manually annotated test collection containing around 6500 English–Chinese pairs.

7.4 Experiments and Discussion

In this section, we describe the experimental results. A total of 61 web sites, which include 26 *.hk* sites and 35 *.cn* sites, were randomly selected from the candidate websites obtained in Section 7.2.1. We crawled about 2.7 GB of web data, comprising approximately 53,000 web pages. We noticed that the quality of the parallel data provided by the *.hk* sites seems to be better than that provided by the *.cn* sites, and therefore we strongly suggest that more importance should be attached to the *.hk* web sites in candidate website selection. We then test the effect of the features, both separately and in various of combinations.

7.4.1 Single feature effect

We run three experiments to separately gauge the effectiveness of each of these features — the file length, the file structure, and the content translation features in RUN_{len}, RUN_{struct}, and RUN_{trans}, respectively. The evaluation results with the highest average precision achieved using tenfold cross-validation are shown in Table 7.2.

Surprisingly, the file length feature, the simplest and thus the most efficient, is clearly superior. When $0.55 \leq \text{S}_{len} < 0.75$, we are able to achieve a precision of 97% and a recall of 70%. This

RUN ID	Precision	Recall
RUN_{len} ($0.55 \leq \text{S}_{len} < 0.75$)	97%	70%
RUN_{struct} ($\text{S}_{struct} \leq 0.1$)	95%	46%
RUN_{trans} ($\text{S}_{trans} \geq 0.1$)	90%	53%
PTMiner	90%	–
STRAND	98%	61%

Table 7.2: Effect of single feature filtering on system effectiveness. For the file length feature, ratios between 0.55 *and* 0.75 *achieved the best precision. For the file structure feature, pairs with scores* ≤ 0.1 *performed best, whereas for the translation feature, attribute scores* ≥ 0.1 *provided the best precision.*

compares favorably to the results of STRAND and PTMiner, which, while not directly comparable because of the the differing corpora, suggests that our system performs reasonably well.

Our utilization of linear sequence of HTML tags to determine whether two pages are parallel, is similar to that of STRAND and PTMiner. The file HTML structure feature provides a relatively high precision; meanwhile, it greatly impairs the recall.

The content translation feature has produced mediocre results. Given Champollion depends heavily on lexical information (previously described in Section 7.2.3), we suspect the main reason is that the majority of the candidate pairs we have generated in Section 7.2.2 are in traditional Chinese, where the bilingual lexicon we have compiled is based on simplified Chinese. Although there are no differences between the basic vocabularies or grammatical structures of simplified and traditional Chinese, different Chinese communities translate English terms in different ways. Due to the limited communication between mainland China (where simplified Chinese is used) and Taiwan, Hong Kong, and the overseas areas (which use traditional Chinese), there are some differences in terminology, especially new cultural or technological nouns.For instance, the English computer phrase 'cross-language information retrieval' is commonly translated in simplified Chinese as "跨语言信息检索", while in traditional Chinese it is "跨語言資訊檢索". This problem was first recognized by Gey et al. [1996] in TREC−5. The common word for 'AIDS' used in Hong Kong and Taiwan is only found in five documents of the TREC−5 simplied Chinese collection. This suggests that better results might be obtained if tailored lexicons were used for mainland and overseas Chinese text.

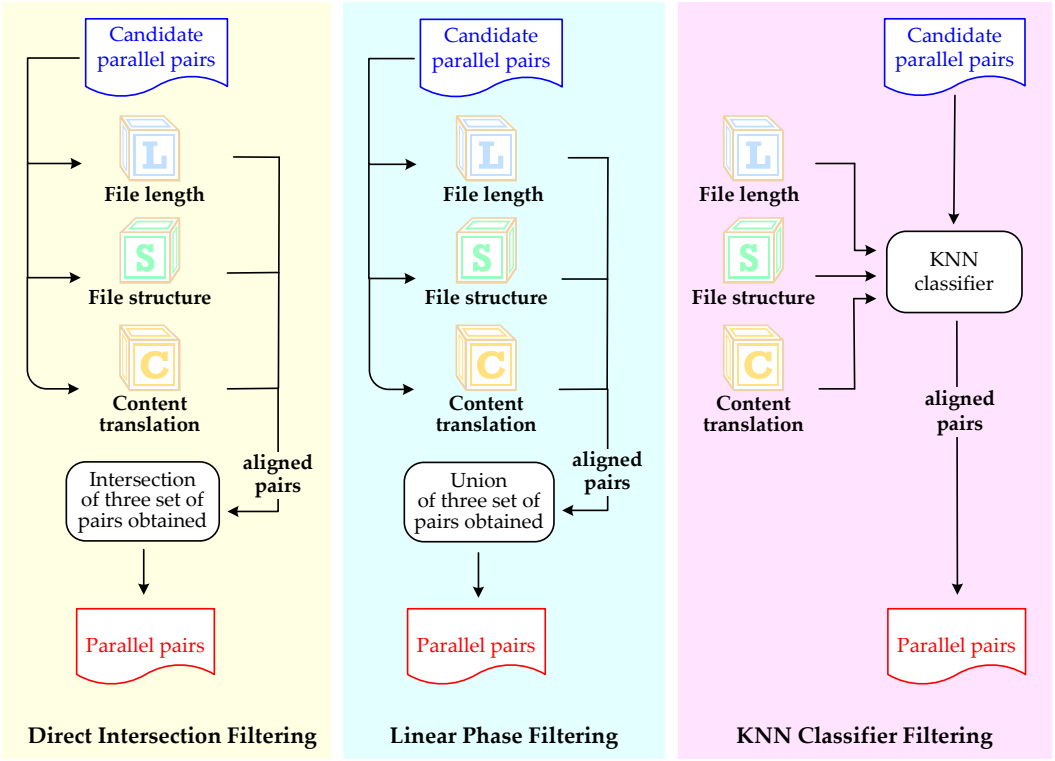


Figure 7.3: Outline of different feature fusion methods: direct intersection filtering, linear phase filtering, and KNN classifier filtering.

7.4.2 Feature fusion effect

The next set of experiments allow us to test whether using feature fusion in the parallel pairs verification is likely to provide any benefit, as well as test the effect of the number of the features of fusion on the overall performance. As shown in Figure 7.3, three types of feature combinations are investigated: direct intersection, linear phase filtering, and a KNN classifier.

In the direct intersection run RUN_{inters}, we evaluate a direct intersection of the pair sets aligned by each of the features. In the linear phase filtering run RUN_{linear}, the candidate pairs are passed through the linear phase filters. The pairs not detected by the first feature filter are aligned using the second feature filter; the pairs left are piped to the last feature filter and processed. In other words, this process produces the union of the sets of pairs aligned by each filter. In the RUN_{knn}, we experiment with a KNN classifier previously described in Section 7.2.4. For example, using a feature space of three dimensions, each pair instance x is represented as a vector $\langle S_{len}(x), S_{struct}(x), S_{trans}(x)\rangle$. RUN_{len} provided the best results for a single feature run,

RUN ID	Features			Precision	Recall
	File length	File structure	Content translation		
RUN_{len} (*Baseline*)	√			97	70
RUN_{inters}	√	√		97	30
	√		√	97	64
		√	√	97	27
	√	√	√	98	20
RUN_{linear}	√	√		95	85
	√		√	95	88
		√	√	94	89
	√	√	√	96	90
RUN_{knn}	√	√		94	94
	√		√	94	97
		√	√	93	97
	√	√	√	95	97

Table 7.3: Effect of various feature fusion filtering on system effectiveness. (All values are percentages.)

and thus is used to establish a reference by which we can measure our feature fusion results. The results reported are obtained after selecting an optimal threshold for each of the feature scores. The experimental results with the highest average precision achieved using tenfold cross-validation are shown in Table 7.3.

The results of the direct intersection combination method (RUN_{inters}) were disastrous. This suggests a large proportion of correct pairs only satisfy some of the above three features. The result of this was often that many correct pairs were omitted. This outcome is corroborated by the results of RUN_{linear}. Using the liner phase filtering feature fusion, we are able to achieve a precision of 96% and a recall of 90%. The KNN classifier further improved the recall to 97%. We used the Wilcoxon signed-rank signed test to test the statistical significance of the improvement. It showed a significant improvement at the 95% confidence level, and emphasizes the importance of a good feature fusion technique.

Our experiments also show that 3-feature fusion statistically significantly outperforms 2-feature fusion in both RUN_{linear} and RUN_{knn}. Therefore we conclude that use of a larger number of features increases the overall performance of the system.

7.5 Summary

In this chapter we described WPED, an automatic mining system for bilingual web parallel corpora. This system used several new techniques to extract parallel web pages, and thus has the potential to find more candidate pages than previous systems.

WPDE uses a three stage process: first, candidate sites are selected and crawled; second, candidate pairs of parallel texts are extracted; finally, we validate the parallel text pairs. Compared to previous systems, WPDE contains improvements at each stage. Specifically, in stage one, in addition to anchor text, image ALT text (the text that always provides a short description of the image and is displayed if an image is not shown) is used to improve the recall of candidate site selction. In stage two, candidate pairs are generated by pattern matching and an edit-distance similarity measure, whereas previous systems only applied one or the other of these. In stage three, where previous systems used a single principle feature to verify parallel pages, WPDE applies a KNN classifier to combine multiple features. Experiments on a large manually annotated data set (containing around 6500 English–Chinese pairs) showed that each of the methods leads to improvements in terms of the overall performance in each step, and that the combined system yielded the best overall result reported.

Our experimental results showed that the use of the KNN classifier with multiple features achieves substantial improvements over the systems that use any one of these features. WPDE has achieved a precision rate of 95% and a recall rate of 97%, and thus is a significant improvement over earlier work.

Chapter 8

Conclusions and Future Work

The Internet has made it possible for people to gain access to more information in various languages than ever before. However, its potential as a global medium is limited by language barriers. Though initially the web was dominated by English speakers, now less than half of existing web pages are in English and only 31.6% of web users are native English speakers, according to the recent research done by Global Reach[1]. This has given rise to a growing interest in obtaining access to information in multiple languages. The purpose of cross-lingual information retrieval (CLIR) is to retrieve relevant documents, where the language of the query is different from that of the documents retrieved. CLIR systems can facilitate routine tasks of groups such as multi-nationals, information services, and government organizations. CLIR research is becoming increasingly valuable for global information exchange and knowledge sharing.

With the rapid growth of the Internet industry in China, its population of Internet users, already the world's second-biggest after the U.S., has increased by nearly 20% over the past year (ending June 2006) to 123 million (China Internet Network Information Center Statistical Survey Report[2], published in July 2006). However, English is the most popular language being used on the web. Chinese web users will inevitably want to be able to access English content using queries expressed in their own scripts. This is driving an increasing need for IR systems that facilitate access to English documents by Chinese web users with varying degrees of expertise with English. CLIR between Western and Asian languages poses significant problems due to the great differences in the structural and written forms of the languages.

In this thesis, we described the difficulties encountered in the Chinese–English web CLIR

[1] www.glreach.com

[2] http://www.cnnic.net.cn/uploadfiles/pdf/2006/7/19/103651.pdf

environment and presented our solutions to these problems. This chapter reviews the primary contributions of the work and then discusses directions for future research.

8.1 Research Contributions

Web queries are often short (with an average 3.18 characters in Chinese [Pu et al., 2002]) and can contain topical terms that relate to current affairs. Short queries provide little context for disambiguation. Topical terms that relate to current affairs are typically absent from static lexicons. Translation ambiguity and OOV terms cause serious problems in dictionary-based query translation in Chinese–English CLIR. In this research, we developed a set of fully automatic techniques to overcome problems with OOV terms and translation ambiguity. A combination of these techniques provided a significant improvement in CLIR effectiveness, and allowed us to achieve up to 97% of monolingual retrieval effectiveness.

Additionally, we developed two automatic web mining systems to extract and build high quality linguistic resources — a lexicon of topical terms and a parallel corpus — for Chinese–English CLIR. In this research, we showed that these two systems are robust, effective, and easy to implement.

This research makes the following contributions to the field of Chinese–English CLIR:

- We proposed a new technique based on a Markov model and the use of web documents as a corpus to provide context for Chinese–English query translation disambiguation. This simple technique has proved to be extremely robust and successful.
- We developed a new segmentation-free technique to identify Chinese OOV terms and extract English translations. This technique does not rely on prior segmentation and is thus free from segmentation error. It leads to a significant improvement in CLIR effectiveness and can also be used to improve Chinese segmentation accuracy.
- Our automatic web mining systems are highly reliable and easy to deploy. In this research, we provided new ways to acquire linguistic resources using multilingual content on the web. These linguistic resources not only improve the efficiency and effectiveness of Chinese–English cross-language web retrieval; but also have wider applications than CLIR.

The following subsections describes each of these contributions in more detail.

Improved translation disambiguation

The translation ambiguity problem is particularly severe in view of the observed tendency of web users to enter short queries; in general it is not possible for even a human to reliably determine

the intended meaning from the available context.

We proposed an improved disambiguation technique for dictionary-based query translation. First, we integrated a contextual window and a distance factor into the language model. Second, we used statistics obtained from the web documents as a corpus, rather than a specific test collection, to provide context for disambiguation. This simple technique has proved to be extremely robust and successful.

Our disambiguation technique is based on a Markov model [Markov, 1971]; such models have been used widely for probabilistic modelling of sequence data. Our principal motivation is that, we believe that two words being in close proximity generally provides stronger correlation and produces more credible results for disambiguation than does co-occurrence of two words in a large window. We therefore investigated the effects of distance factor and window size when using a Markov model to provide disambiguation. Contrary to what has been noted when using mutual information techniques to provide disambiguation, we observed that using a window distance factor has no benefit when combined with a Markov model. This showed that the Markov model is a superior technique where sequence data is involved, and is not significantly improved by the addition of a distance factor, thus potentially reducing the computational cost.

In previous work, researchers have relied on the test collection corpus to perform translation disambiguation in CLIR. Using a pre-defined training corpus for disambiguation has limitations in both availability and coverage. Since the web consists of documents in various domains or genres, we investigated how to utilize the web documents as a corpus to provide context for translation disambiguation. Our experimental results showed that, when using the web, it is possible to achieve effectiveness comparable to that obtained with a pre-defined training corpus.

We also explored other alternatives for translation disambiguation. To understand the relative merits of the different disambiguation techniques, we compare our technique to other approaches using the same Chinese–English CLIR test collections. Our results showed that despite the different underlying models and formulae used, the aggregated results are comparable. However, there is wide variation in the translation of individual queries, suggesting that there is scope for further improvement.

Segmentation-free OOV term translation

OOV terms are generally special vocabulary — such as technical terms, place names, brand names, or personal names — that are outside the scope of translation dictionaries. An OOV solution would be useful for a wide range of tasks, including CLIR, lexical acquisition, and mobile personal IR,

where a person may be unfamiliar with the correct foreign language terms or expression to use in a specific context.

We developed a novel technique to dynamically discover translations of OOV terms through the mining of web text, based on the Chinese authorial practice of including unfamiliar English terms in both languages. Our technique does not rely on prior segmentation and is thus free from segmentation error. It can correctly extract Chinese–English translations that were previously undetected, or only detected after manual intervention to provide correct segmentation. Additionally, the extracted Chinese terms can be used to enhance a Chinese segmentation dictionary and thus improve Chinese segmentation accuracy.

In experiments with our OOV term translation technique using several collections and a set of terms from news articles, our results showed that this technique is robust and provides a substantial improvement in CLIR effectiveness. This technique can be applied to both Chinese–English and English–Chinese CLIR, correctly extracting translations of OOV terms from the web automatically, and thus is a significant improvement on earlier work.

Automatic linguistic resources acquisition

Linguistic resources such as bilingual lexicons and parallel web text are essential in CLIR. The effectiveness of a CLIR system is inevitably limited by the caliber of translations; clearly resources with wide coverage and extensive information are preferable. However, high quality linguistic resources are typically difficult to obtain and exploit, or expensive to purchase. The amount of multilingual information available on the web is expanding rapidly, and provides a valuable new set of linguistic resources. In this research, we developed two automatic mining systems to make use of publicly available translation resources on the web:

- AutoLex, a system to extract topical translations from Chinese text on the web and automatically update Chinese–English (as well as English–Chinese) translation lexicons via periodic crawls of the web;
- Web Parallel Data Extraction (WPDE), a system to automatically collect high quality Chinese–English parallel corpora from the web.

In this work, we have shown that these two system are highly reliable and easy to deploy.

AutoLex — OOV term lexicon construction system

We developed an automatic system (AutoLex) to facilitate construction of a large-scale translation lexicon of OOV terms using the web. AutoLex extracts topical translations from Chinese text on the web and automatically updates Chinese–English (as well as English–Chinese) translation lexicons via periodic crawls of the web.

Inconsistencies in how the translations are presented lead to ambiguity; however, we showed that these can be resolved statistically. Of the Chinese–English text snippets we collected, the great majority were cases where the Chinese text was not punctuated, and thus there was no indication of where the translation began and finished; and most of these only occur once. The use of syntactical structure has allowed us to greatly increase the accuracy of translation in circumstances where the co-occurrence frequency of translation pairs is low, and thus greatly increase the size of automatically-created dictionary.

We used a native Chinese speaker (not otherwise associated with this research) to evaluate the translations we extracted from the web. From 89,848 extracted OOV translation pairs, 400 pairs were randomly selected, giving an error margin of ±5%. As shown in Section 6.4.1, 63% of the extracted translation pairs were strictly correct. However, an additional 17% of translation pairs contained the correct translation, with some additional related information, such as the person's title or the organization's location. When used for retrieval, this additional information may arguably improve effectiveness. Such translation pairs may be added to both English-to-Chinese and Chinese-to-English lexicons, that can later be used at query time, thus improving the efficiency of query processing.

Moreover, such a system has wider applications than CLIR; for example, discovered translations of OOV terms can guide development of printed translation dictionaries and can be used in machine translation of full text.

WPDE — Web Parallel Data Extraction System

In order to take advantage of publicly available parallel corpora, we developed WPDE system, that combines multiple features to identify parallel texts via a k-nearest-neighbor (KNN) classifier, to automatically collect high quality parallel Chinese–English corpora from the web.

Compared to previous systems, WPDE used several improved techniques to extract parallel web pages. First, in addition to anchor text, image ALT text is used to improve the recall of candidate site selection. Second, candidate pairs are generated by pattern matching and an edit-distance similarity measure, whereas previous systems only applied one or the other of these.

Third, where previous systems used a single principle feature to verify parallel pages, WPDE applies a KNN classifier to combine multiple features.

Experiments on a large manually annotated data set (containing around 6500 English–Chinese pairs) showed that each of the new techniques led to improvements in terms of the overall performance, and that the combined system yielded the best overall results. Our results (as described in Section 7.4.2) also showed that the use of the KNN classifier with multiple features achieves substantial improvements over the systems that use any one of these features. WPDE has achieved a precision rate of 95% and a recall rate of 97%, and is a significant improvement over earlier work.

8.2 Future Work

Building on our research and system development experiences, we raise the following issues in CLIR for further investigation:

- How to automatically detect the unreliable translations in a translation dictionary?
- How can different disambiguation techniques be combined to improve the overall translation quality in CLIR?
- What are the most appropriate forms of query expansion and document expansion for use in Chinese–English CLIR?

Automatic verification of translation dictionary

Dictionary coverage is an important issue for CLIR systems. In this thesis, we designed and developed a method to automatically discover the translations of new vocabulary through mining web text and thus alleviating the OOV problem common in dictionary-based CLIR.

Dictionary quality is the other important factor affecting the performance of CLIR systems. The unreliable translation problem occurs because there is no simple way to detect the presence of unreliable translations. Sometimes the dictionary only provides wrong or bad translations. Therefore, we can fail to select a correct translation using our disambiguation technique because there is no correct translation for the query as a whole, even though each query key term has at least one translation. In Section 3.3.4, we showed a comparison of the interpolated recall-precision averages for the English–Chinese CLIR experimental results. The performance of our query translation system was degraded by the unreliable translation problem. Some translations provided by the translation dictionary are inappropriate in the given context; for example, in

CH28, the English term "cellular phone" is translated into "汽车电话" (car phone), where the given Chinese equivalent is "移动电话". Additionally, in CH47, "impact" is translated into "冲击" (strike), where the given Chinese equivalent is "后果".

The quality of the translation resources has a significant impact on CLIR effectiveness. To develop statistical techniques for automatic dictionary verification will allow us to detect and correct unreliable translations, and thus improve CLIR performance.

Using combination of evidence for translation disambiguation

Translation ambiguity is a frequent cause of failure of dictionary-based translation, because many words or phrases in one language can be translated into another language in multiple ways, and sometimes the alternate translations have very different meanings. In this thesis, we developed techniques based on a Markov model and the utilization of web documents as a corpus to provide context for translation disambiguation. We also explored other alternatives for translation disambiguation. In order to understand the relative merits of the different disambiguation techniques on the basis of different underlying principles, in Section 4.3.3, we compared our technique to other approaches on the same Chinese–English data sets.

Our results showed that, superficially, all these translation disambiguation methods are comparable (no significant difference) when averaged across a query set. However, at the individual query level, each of the techniques frequently produce differing results. This means that it may be possible to develop a new approach that combines the best elements of these current methods. At the very least, it should be possible to use a "combination of evidence" approach [Smets, 1990] to improve the overall translation quality. Further, when all methods produce the same translation, we can have a higher degree of confidence in the correctness, and differing translations could act as a trigger for further analysis.

It has been recognized that the combination of different sources of evidence can improve the effectiveness of IR [Croft, 2000]. In dictionary-based query translation, assuming that not all queries benefit from the same disambiguation method, the most appropriate disambiguation method should be determined for a specific query type. A selection mechanism should be developed for associating the optimal disambiguation method to each type of queries. For example, for type A queries we can employ any one of these available disambiguation methods; while for type B queries, we use evidence from the term similarity method and the mutual information method. By combining the results of several independent translation disambiguation techniques, it is possible to obtain performance better than that of the best individual disambiguation technique.

Combined document and query expansion techniques

In nearly all cases, CLIR leads to some loss of retrieval effectiveness due to inexact language mappings. Query expansion has long been suggested as an effective way to resolve the short query and word mismatching problems in monolingual retrieval [Mitra et al., 1998].

The aim of query expansion is to reduce query-document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. In this thesis, we evaluated a two-stage procedure of selecting additional query terms for post-translation query expansion based on term weighting and word association information. This technique provided improvements of 11% for Chinese–English CLIR effectiveness. Comparable document-side expansion is a relatively more recent development motivated by error-prone transcription and translation processes in spoken documents and CLIR. Document expansion enriches the documents with highly selective terms drawn from highly ranked documents retrieved by using the document itself as a query [Singhal and Pereira, 1999].

Combined expansion techniques, both query expansion and document expansion, using the document collection should be particularly useful and effective in Chinese–English CLIR.

Appendix A

56 English seed terms

Adnan Pachachi
Aerosmith
Ahmad Qurei
Akitsugu Konno
Alan Jackson
Allen Moo
Amel Larrieux
Bonnie Raitt
Brian McKnight
CELESTICA
Chad Hugo
Charlize Theron
Chevron Texaco
Chiharu Igaya
Christian Laettner
CNNIC
Cynthia Cooper
Dan Churchaid
Darrell Hammond
Gracie Mansion
GUCCI
Hercules Inlet
Jacob Zuma
Jerusalem
Junichiro Koizumi
KrollAssociates
Kurata Hiroyuki
LCD
Mike Zafirovski
Moshe Katsav
muhammed hosni mubarak
MURCIELAGO
NASA
Neptune
Newt Gingerich
Nirvana
NIST
Nizar Khazraji
OBILIC
OPEC
OSDL
PricewaterhouseCoopers
RENEE ZELLWEGER
Saab Military Aircraft
Shaul Mofaz
SOFIA COPPOLA
Sofitel
Stanley Clarke
Tarantula Nebula
Tikrit
Troyen Brennan
Tulane University
UN University
Watanuki Tamisuke
Yoriko Kawaguchi
Zendik

Appendix B

A list of pre-defined strings

english
chinese
simplifiedchinese
chinesesimplified
traditionalchinese
chinesetraditional
englishversion
simplifiedchineseversion
traditionalchineseversion
英文
简体
繁體
英文版
中文版
简体版
繁體版
英文网站
中文网站
英文首页
中文首页
中文简体
中文繁體
简体中文
简体中文版
繁體中文
繁體中文版

Bibliography

M. Adriani. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information Retrieval*, 2(1):71–82, 2000.

M. Aljlayl and O. Frieder. Effective Arabic–English cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 295–302, Atlanta, Georgia, USA, 2001. ACM Press.

J. Allan, M. E. Connell, W. B. Croft, F.-F. Feng, D. Fisher, and X. Li. INQUERY and TREC-9. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 551–562, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

L. Ballesteros. Cross-language retrieval via transitive translation. In W. B. Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 203–234. Kluwer Academic Publishers, 2000.

L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, Philadelphia, PA, USA, 1997. ACM Press.

L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, Melbourne, Australia, 1998. ACM Press.

B. Billerbeck and J. Zobel. Questioning query expansion: an examination of behavior and parameters. In K.-D. Schewe and H. E. Williams, editors, *Proceedings of the 15th Australasian*

Database Conference, volume 27, pages 69–76, Dunedin, New Zealand, 2004. Australian Computer Society, Inc.

M. Braschler and C. Peters. CLEF 2003 methodology and metrics. In Peters et al. [2004], pages 7–20.

M. Braschler, C. Peters, and P. Schäuble. Cross-language information retrieval (CLIR) track overview. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*, pages 25–34, Gaithersburg, Maryland, 1999. Department of Commerce, National Institute of Standards and Technology.

M. Braschler, G. M. D. Nunzio, N. Ferro, and C. Peters. CLEF 2004: Ad hoc track overview and results analysis. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 10–26. Springer, 2004.

M. R. Brent and X. Tao. Chinese text segmentation with MBDP-1: making the most of training corpora. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 90–97, Toulouse, France, 2001. Association for Computational Linguistics.

E. Brill. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79–85, 1990.

R. D. Brown. Example-based machine translation in the pangloss system. In *Proceedings of the 16th conference on Computational Linguistics*, pages 169–174, Copenhagen, Denmark, 1996. Association for Computational Linguistics.

R. Chau and C.-H. Yeh. Construction of a fuzzy multilingual thesaurus and its application to cross-lingual text retrieval. In *Proceedings of the 1st Asia-Pacific Conference on Web Intelligence: Research and Development*, pages 340–345, Maebashi City, Japan, 2001. Springer-Verlag.

A. Chen and F. Gey. Experiments on cross-language and patent retrieval at NTCIR-3 workshop. In *Proceedings of the 3rd NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2003. National Institute of Informatics, Japan.

A. Chen and F. C. Gey. Combining query translation and document translation in cross-language retrieval. In Peters et al. [2004].

A. Chen, J. He, L. Xu, F. C. Gey, and J. Meggs. Chinese text retrieval without using a dictionary. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, Philadelphia, Pennsylvania, USA, 1997. ACM Press.

A. Chen, H. Jiang, and F. Gey. Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 17–23, Hong Kong, China, 2000. ACM Press.

H.-H. Chen, C.-C. Lin, and W.-C. Lin. Building a Chinese-English wordnet for translingual applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(2): 103–122, 2002.

J. Chen and J.-Y. Nie. Parallel web text mining for cross-language IR. In *Proceedings of the 6th RIAO Conference on Content-Based Multimedia Information Access*, pages 62–77. College de France, 2000.

J. Chen, R. Chau, and C.-H. Yeh. Discovering parallel text from the world wide web. In *Proceedings of the 2nd Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, pages 157–161, Dunedin, New Zealand, 2004. Australian Computer Society, Inc.

P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153, Sheffield, UK, 2004. ACM Press.

L.-F. Chien. PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–58, Philadelphia, PA, USA, 1997. ACM Press.

K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas, USA, 1988. Association for Computational Linguistics.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Journal of the Computational Linguistics*, 16(1):22–29, 1990.

C. Cleverdon. The Cranfield tests on index language devices. *Readings in Information Retrieval*, pages 47–59, 1997.

C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, Illinois, USA, 1991. ACM Press.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

W. B. Croft. Combining approaches in information retrieval. *Advances in Information Retrieval: Recent Research from the CIIR*, pages 1–36, 2000.

Y. Dai, T. E. Loh, and C. S. G. Khoo. A new statistical formula for Chinese text segmentation incorporating contextual information. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 82–89, Berkeley, CA, USA, 1999. ACM Press.

D. Farwell, L. Guthrie, and Y. Wilks. The automatic creation of lexical entries for a multilingual MT system. In *Proceedings of the 14th Conference on Computational linguistics*, pages 532–538, Nantes, France, 1992. Association for Computational Linguistics.

M. Federico and N. Bertoldi. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 167–174, Tampere, Finland, 2002. ACM Press.

F. Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.

M. Franz, J. S. McCarley, T. Ward, and W.-J. Zhu. Quantifying the utility of parallel corpora. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 398–399, New Orleans, Louisiana, USA, 2001. ACM Press.

A. Fujii and T. Ishikawa. Applying machine translation to two-stage cross-language information retrieval. In J. S. White, editor, *AMTA*, volume 1934 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2000.

J. Gao, J.-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou, and C. Huang. TREC-9 CLIR Experiments at MSRCN. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 343–378, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–104, New Orleans, Louisiana, USA, 2001. ACM Press.

J. Gao, M. Zhou, J.-Y. Nie, H. He, and W. Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 183–190, Tampere, Finland, 2002. ACM Press.

X. Ge, W. Pratt, and P. Smyth. Discovering Chinese words from unsegmented text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–272, Berkeley, CA, USA, 1999. ACM Press.

F. C. Gey and A. Chen. TREC-9 cross-language information retrieval (English–Chinese). In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 15–24, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

F. C. Gey and H. Jiang. English–German cross-language retrieval for the GIRT collection — exploiting a multilingual thesaurus. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*, pages 301–306, Gaithersburg, Maryland, 1999. Department of Commerce, National Institute of Standards and Technology.

F. C. Gey, A. Chen, J. He, L. Xu, and J. Meggs. Term importance, boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5. In E. M.

Voorhees and D. K. Harman, editors, *NIST Special Publication 500-238: The Fifth Text REtrieval Conference (TREC-5)*, pages 181–190, Gaithersburg, Maryland, 1996. Department of Commerce, National Institute of Standards and Technology.

F. C. Gey, N. Kando, and C. Peters. Cross-language information retrieval: the way ahead. *Information Processing and Management*, 41(3):415–431, 2005.

T. Gollins and M. Sanderson. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 90–95, New Orleans, Louisiana, USA, 2001. ACM Press.

D. Hiemstra and F. de Jong. Disambiguation strategies for cross-language information retrieval. In *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, pages 274–293, Paris, France, 1999. Springer-Verlag.

D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *Cross-Language Information Retrieval and Evaluation*, pages 102–115, Berlin, 2001. Springer-Verlag.

F. Huang, S. Vogel, and A. Waibel. Extracting named entity translingual equivalence with limited resources. *ACM Transactions on Asian Language Information Processing*, 2(2):124–129, 2003.

D. A. Hull and G. Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, Zurich, Switzerland, 1996. ACM Press.

M.-G. Jang, S. H. Myaeng, and S. Y. Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 223–229, College Park, Maryland, 1999. Association for Computational Linguistics.

H. Jin and K.-F. Wong. TREC-9 CLIR at CUHK disambiguation by similarity values between adjacent words. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 151–188, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

W. Jin. Chinese segmentation disambiguation. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1245–1249, Kyoto, Japan, 1994. Association for Computational Linguistics.

K. Kishida. Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41(3):433–455, 2005.

K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, S. H. Myaeng, and K. Eguchi. Overview of CLIR task at the fourth NTCIR workshop. In *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004. National Institute of Informatics, Japan.

K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of CLIR task at the fifth NTCIR workshop. In *Proceedings of the 5th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2005. National Institute of Informatics, Japan.

W. Kraaij. TNO at CLEF-2001: comparing translation resources. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *CLEF*, volume 2406 of *Lecture Notes in Computer Science*, pages 78–93. Springer, 2001.

W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Journal of the Computational Linguistics*, 29(3):381–419, 2003.

K.-L. Kwok. Exploiting a Chinese–English bilingual wordlist for English–Chinese cross-language information retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 173–179, Hong Kong, China, 2000. ACM Press.

K.-L. Kwok, L. Grunfeld, N. Dinstl, and M. Chan. TREC-9 cross language, web and question-answering track experiments using PIRCS. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 419–436, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

K.-L. Kwok, N. Dinstl, and S. Choi. NTCIR-4 Chinese, English, Korean cross language retrieval experiments using PIRCS. In *Proceedings of the 4th NTCIR Workshop on Research in Infor-*

mation Access Technologies: Information Retrieval, Question Answering and Summarization, pages 186–192, Tokyo, Japan, 2004. National Institute of Informatics, Japan.

G.-A. Levow. University of Chicago at NTCIR4 CLIR: multi-scale query expansion. In *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, pages 78–85, Tokyo, Japan, 2004. National Institute of Informatics, Japan.

H.-Z. Li and B.-S. Yuan. Chinese word segmentation. In *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation*, pages 212–217, Singapore, 1998.

W.-H. Lin and H.-H. Chen. Backward machine transliteration by learning phonetic similarity. In *Proceedings of the 6th Conference on Computational Natural Language Learning*, pages 139–145, Taipei, Taiwan, 2002. Morgan Kaufman Publishers.

W.-H. Lu, L.-F. Chien, and H.-J. Lee. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing*, 1(2):159–172, 2002.

A. Maeda, F. Sadat, M. Yoshikawa, and S. Uemura. Query term disambiguation for web cross-language information retrieval using a search engine. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 25–32, Hong Kong, China, 2000. ACM Press.

A. A. Markov. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. *reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems: Markov Models*, 1, 1971.

J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214, College Park, Maryland, 1999. Association for Computational Linguistics.

C. J. A. McEwan, I. Ounis, and I. Ruthven. Building bilingual dictionaries from parallel web documents. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 303–323. Springer-Verlag, 2002.

P. McNamee and J. Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval, pages 159–166, Tampere, Finland, 2002. ACM Press.

H. M. Meng, B. Chen, S. Khudanpur, G.-A. Levow, W.-K. Lo, D. W. Oard, P. Schone, K. Tang, H.-M. Wang, and J. Wang. Mandarin–English information (MEI): investigating translingual speech retrieval. *Computer Speech and Language*, 18(2):163–179, 2004.

M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, Melbourne, Australia, 1998. ACM Press.

C. Monz and B. J. Dorr. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 520–527, Salvador, Brazil, 2005. ACM Press.

H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(3):1–38, 1994.

J.-Y. Nie. TREC-7 CLIR using a probabilistic translation model. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC-7)*, pages 547–554, Gaithersburg, Maryland, 1998. Department of Commerce, National Institute of Standards and Technology.

J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, Berkeley, CA, USA, 1999. ACM Press.

D. W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 472–483, London, UK, 1998. Springer-Verlag.

D. W. Oard. Evaluating interactive cross-language information retrieval: document selection. In *Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, pages 57–71, London, UK, 2001. Springer-Verlag.

D. W. Oard and A. R. Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–256, 1998.

P. Pantel, A. Philpot, and E. Hovy. Aligning database columns using mutual information. In *Proceedings of the 2005 National Conference on Digital Government Research*, pages 205–210, Atlanta, Georgia, 2005. Digital Government Research Center.

F. Peng and D. Schuurmans. Self-Supervised Chinese word segmentation. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, pages 238–247, Cascais, Portugal, 2001. Springer-Verlag.

C. Peters and P. Sheridan. Multilingual information access. pages 51–80, 2001.

C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors. *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers*, volume 3237 of *Lecture Notes in Computer Science*, 2004. Springer.

A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, Melbourne, Australia, 1998. ACM Press.

M. F. Porter. An algorithm for su#x stripping. *Automated Library and Information Systems*, 14 (3):130–137, 1980.

H.-T. Pu, S.-L. Chuang, and C. Yang. Subject categorization of query terms for exploring web users' search interests. *Journal of the American Society for Information Science and Technology*, 53(8):617–630, 2002.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267–296, 1990.

F. Ren. A hybrid approach of text segmentation based on sensitive word concept for NLP. In *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing*, pages 375–388, Mexico City, Mexico, 2001. Springer-Verlag.

P. Resnik and N. A. Smith. The web as a parallel corpus. *Journal of the Computational Linguistics*, 29(3):349–380, 2003.

C. J. V. Rijsbergen. *Information Retrieval*. London: Butterworths, 2nd edition, 1979.

A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, Berkeley, California, USA, 1999. ACM Press.

P. Smets. The combination of evidence in the transferable belief model. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3): 226–234, 2001.

R. Sproat and C. Shih. A statistical method for finding word boundaries in Chinese text. *Journal of the Computer Processing of Chinese and Oriental Languages*, 4(4):336–351, 1990.

J. Sun, M. Zhou, and J. Gao. A class-based language model approach to Chinese named entity identification. *Computational Linguistics and Chinese Language Processing*, 8(2):1–28, 2003.

W. J. Teahan, R. McNab, Y. Wen, and I. H. Witten. A compression-based algorithm for chinese word segmentation. *Computational Linguistics*, 26(3):375–393, 2000.

E. M. Voorhees. Overview of TREC 2002. In E. M. Voorhees and L. P. Buckland, editors, *NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)*, Gaithersburg, Maryland, 2002a. Department of Commerce, National Institute of Standards and Technology.

E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, (Revised Papers)*, pages 355–370, London, UK, 2002b. Springer-Verlag.

R. A. Wagner and R. Lowrance. An extension of the string-to-string correction problem. *Journal of the ACM*, 22(2):177–183, 1975.

G. Wang and F. H. Lochovsky. Feature selection with conditional mutual information maximin in text categorization. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 342–349, New York, NY, USA, 2004. ACM Press.

J.-H. Wang, J.-W. Teng, P.-J. Cheng, W.-H. Lu, and L.-F. Chien. Translating unknown cross-lingual queries in digital libraries using a web-based approach. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 108–116, Tucson, Arizona, USA, 2004. ACM Press.

D. R. Wilson and T. R. Martinez. Instance pruning techniques. In *Proceedings of the 14th International Conference on Machine Learning*, pages 403–411, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, Los Altos, CA, USA, 2nd edition, 1999.

P.-K. Wong and C. Chan. Chinese word segmentation based on maximum matching and word binding force. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 200–203, Copenhagen, Denmark, 1996. Association for Computational Linguistics.

L. Wu, X.-J. Huang, Y. Guo, B. Liu, and Y. Zhang. FDU at TREC-9: CLIR, filtering and QA tasks. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 189–222, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

J. Xu and R. Weischedel. TREC-9 Cross Lingual Retrieval at BBN. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 106–150, Gaithersburg, Maryland, 2000. Department of Commerce, National Institute of Standards and Technology.

J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–110, New Orleans, Louisiana, USA, 2001. ACM Press.

E. Xun, C. Huang, and M. Zhou. A unified statistical model for the identification of English baseNP. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 109–116, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

C. C. Yang and K. W. Li. A heurisitic method based on a statistical approach for Chinese text segmentation: Research Articles. *Journal of the American Society for Information Science and Technology*, 56(13):1438–1447, 2005.

C. C. Yang and K. W. Li. Mining English/Chinese parallel documents from the world wide web. In *Proceedings of the International World Wide Web Conference*, pages 188–192. Honolulu, Hawaii, 2002.

C. C. Yang, J. W. K. Luk, S. K. Yung, and J. Yen. Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science*, 51(4): 340–351, 2000.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

J. Zhang, J. Gao, and M. Zhou. Extraction of Chinese compound words: an experimental study on a very large corpus. In *Proceedings of the 2nd workshop on Chinese Language Processing*, pages 132–139, Hong Kong, China, 2000. Association for Computational Linguistics.

Y. Zhang and P. Vines. Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pages 162–169, Sheffield, UK, 2004a. ACM Press.

Y. Zhang and P. Vines. Detection and translation of OOV terms prior to query time. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pages 524–525. ACM Press, 2004b.

Y. Zhang and P. Vines. RMIT Chinese–English CLIR at NTCIR-4. In *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, pages 60–64, Tokyo, Japan, 2004c. National Institute of Informatics, Japan.

Y. Zhang and P. Vines. Using the web for translation disambiguation (RMIT university at NTCIR-5 Chinese–English CLIR). In *Proceedings of the 5th NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, pages 69–74, Tokyo, Japan, 2005. National Institute of Informatics, Japan.

Y. Zhang and P. Vines. Improved cross-language information retrieval via disambiguation and vocabulary discovery. In *Proceedings of the 8th Australasian Document Computing Symposium*, pages 3–7, Canberra, Australia, 2003. CSIRO ICT Centre.

Y. Zhang, P. Vines, and J. Zobel. Chinese OOV translation and post-translation query expansion in Chinese–English cross-lingual information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):57–77, 2005.

Y. Zhang, P. Vines, and J. Zobel. An empirical comparison of translation disambiguation techniques for Chinese–English cross-language information retrieval. In *Proceedings of the 3rd Asia Information Retrieval Symposium*, 2006a.

Y. Zhang, K. Wu, J. Gao, and P. Vines. Automatic acquisition of Chinese–English parallel corpus from the web. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *ECIR*, volume 3936 of *Lecture Notes in Computer Science*, pages 420–431. Springer, 2006b.